The Detection of Ischemic Stroke on the PSD Manifold of EEG Signals

By Canxiu Zhang ,

B.Eng.

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University

Master of Applied Science (2018)

Hamilton, Ontario (Department of Electrical and Computer Engineering)

TITLE: The Detection of Ischemic Stroke on the PSD Manifold of EEG Signals

AUTHOR: Canxiu Zhang (McMaster University)

SUPERVISOR: Dr. Kon Max Wong

NUMBER OF PAGES: xi, 77

Abstract

The study of ischemic brain stroke detection by *Electroencephalography* (*EEG*) signal is the area of binary signal classification. In general, this involves extracting features from *EEG* signal on which the classification is performed. In this thesis, we investigate the employment of *Power Spectral Density* (*PSD*) matrix, which contains not only power spectrum contents of each signal which complies with what clinical experts use in their visual judgement of *EEG* signals, but also cross-correlation between multi-channel (electrodes) signals to be studied, as a feature in signal classification. Since the *PSD* matrices are structurally constrained, they form a *manifold* in the signal space. Thus, the commonly used Euclidean distance to measure the similarity/dissimilarity between two *PSD* matrices are not informative or accurate. *Riemannian Distance* (*RD*), which measures distance along the surface of the manifold, should be employed to give more meaningful measurements. Furthermore, two classification methods, binary hypothesis testing and *K-Nearest Neighbors* (*KNN*), are applied. In order to enhance the detection performance, algorithms to find optimum weighting matrix for each classifier are also applied. Experimental results show that the performance by the kNN method using *PSD* matrix as features with *RD* as similarity/dissimilarity measurements are very encouraging.

Acknowledgements

I wish to express my gratitude to all those people who gave me the possibility to complete this thesis. First and foremost, I would like to express the deep gratitude to my supervisor Dr. Kon Max Wong, who offered me the enthusiastic support, patience, encouragement and guidance in my research over the past two years. I would like to thank my co-supervisor Dr. JianKang Zhang for helping discussions, suggestions, and guidance during my study. In addition, I would like to thank Dr. Rong Zheng for patience and valuable suggestions throughout my study and research.

Furthermore, I would like to appreciate all the colleagues and staff members at McMaster University who have provided a supportive and exciting research atmosphere.

Finally, I want to express gratitude to my parents and friends for their continuous supports.

Contents

List of Figures

List of Tables

Acronyms

ACF *Autocorrelation Function*

ANN *Artificial Neuron Network*

AUC *Area Under the Receiver Operating Characteristic Curve*

EEG *Electroencephalography*

ELM *Extreme Learning Machine*

KNN *K-Nearest Neighbors*

PSD *Power Spectral Density*

RD *Riemannian Distance*

ROC *Receiver Operating Characteristic*

VAR *Vector Auto-Regression*

WSS *Wide Sense Stationary*

List of Symbols

$(\cdot)^H$	Matrix hermitian
$(\cdot)^T$	Matrix transpose
$(\cdot)^{-1}$	Matrix inverse
$\langle \cdot, \cdot \rangle$	Inner product
$\mathbb{E}(\cdot)$	Expectation
$\mathcal{H}$	Euclidean space
$\mathcal{M}$	Manifold
$\mathcal{R}(\cdot)$	Real part of a matrix
$\mathbf{0}$	Zero matrix
$\mathbf{I}_M$	$M \times Midentitymatrix$
$\mathbf{S}$	Matrices
$\mathbf{s}$	Column vectors
$\lvert \cdot \rvert$	Magnitude of a complex quantity
$\lvert\lvert \cdot \rvert\rvert$	Euclidean norm of a vector or a matrix
$diag\{\cdot\}$	Diagonal matrix
$tr(\cdot)$	Trace of matrices
$vec\{\cdot\}$	Vectorization of a matrix

Chapter 1

Introduction

1.1 Introduction

Stroke is the second leading cause of death worldwide [Mozaffarian 2015, (WHO) et al. n.d.]. One out of twenty deaths in America are due to stroke and 62,000 strokes that occur each year in Canada affect all age groups and lead to a lifetime impact on health [Mozaffarian 2015]. According to centers of disease control and prevention, in the United States, someone suffers from brain stroke in every 40 seconds.

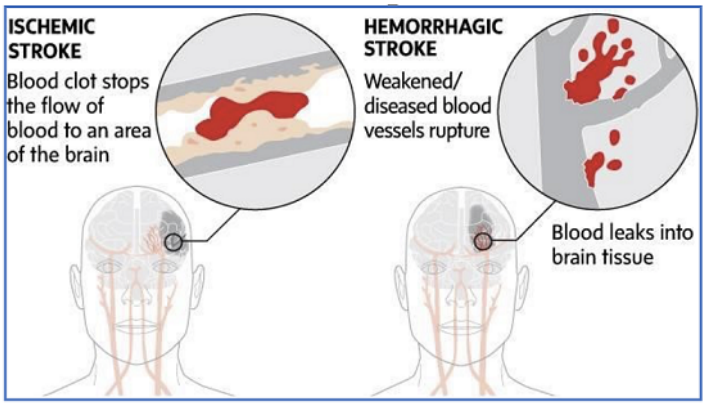

FIGURE 1.1: *Types of Strokes: Ischemic and Hemorrhagic. In Ischemic Brain Stroke (left), a blood clot has blocked the flow of blood to a specific area of the brain.*

There are two main types of stroke: ischemic and hemorrhagic. Ischemic stroke occurs when a blockage (obstruction) of small blood vessels occurs around the brain (Fig.1.1). Magnet Resonance Imaging (MRI) gives accurate results for stroke detection, yet it is expensive, requires several hours to generate an examination report and is applicable for a limited time. MRI is used only in situations where there is no time pressure to offer diagnosis, typically as follow-up imaging. MRI is an expensive and is not available at all healthcare locations. In comparison, *Electroencephalography* (*EEG*) offers a continuous, real-time, non-invasive measure of brain function [Foreman and Claassen 2012] and is capable of detecting ischemic stroke due to variation in cerebral blood flow in the blood vessels. It has proven to be effective in detecting various other brain-related activities like Rapid Eye Movement (REM), sleep and awake stage and other seizure [Röschke and Aldenhoff 1992].

1.1.1 Background knowledge of *EEG* signals for brain stroke detection

EEG signals serve as a vital source of information when it comes to brain function. It is possible to recognize abnormal activities of the brain functionality using *EEG* signals. Most of the cerebral signal observed in scalp falls in the range of 1–20 Hz. The majority of the *EEG* used in clinical practice subdivides the waveforms into bandwidths known as alpha, beta, theta, and delta (Fig.1.2) [Foreman and Claassen 2012].

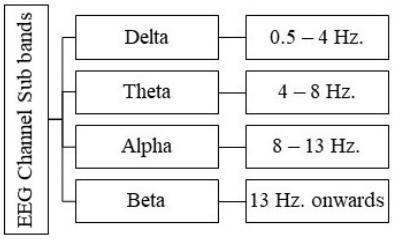

FIGURE 1.2: *EEG Channel Sub bands Frequency range.*

An ischemic stroke is primarily due to changes in Cerebral Blood Flow (CBF), and it can be detected through changes in *EEG* signal patterns [Kantelhardt et al. 2002]. Prominent changes in CBF include the reduction of delta (lowest frequency band) or

the presence of high-frequency bands (beta and alpha) [Foreman and Claassen 2012]. Furthermore, the power density ratio between bands of the different hemisphere changes as stroke affects one hemisphere [Kantelhardt et al. 2002]. The best results for Ischemic stroke detection so far, are obtained using MRI scans in conjunction with meta-data like patients' history, medical prescriptions and most importantly MRI scan [(Tang et al. 2017)].

[Omar et al. 2014a] has implemented *Power Spectral Density* (*PSD*) ratio between different *EEG* bands and channels, as features for ischemic stroke detection. In his research, he has used *PSD* to train *K-Nearest Neighbors* (*KNN*), *Artificial Neuron Network* (*ANN*) [(Tang et al. 2017)] and *Extreme Learning Machine* (*ELM*).

In *EEG* signal analysis process, feature extraction is done by utilizing series of transformations so that the required information can be studied or observed easily in the transform domain to provide the best input to the classifier [Motomura et al. 2015].

1.1.2 Contributions

In this thesis, we investigate the problem of detection of ischemic stroke using *EEG* signals of the patients. Due to the rich correlation (second-order) information contained in the *PSD* matrix of the *EEG* signal, we also employ it as the signal *feature* to facilitate the stroke detection.

However, to directly employ the *PSD* matrices in signal processing, we often have to measure the distance between these features. Being positive semi-definite and Hermitian symmetric, the *PSD* matrices are structurally constrained and thus form a *manifold* $\mathcal{M}$ in the *real* linear vector space $\mathcal{H}$ of all $M \times M$ matrices [Li and Wong 2013]. Therefore, the commonly used Euclidean distance (ED) may not be appropriate for measuring the distance between two *PSD* matrices; rather, we should measure the distance along the surface of the manifold. This concept is akin to finding the distance between two cities on earth: The ED between two cities is neither informative nor accurate. By the same reasoning, we realize that the distance between two of these matrices is more accurately

measured along the surface of the *PSD* manifold, i.e., by the *Riemannian distance* (RD). In particular, we employ a RD d_{R_2}, together with its weighted version d_{WR_2}, which are suitable for signal processing. Furthermore, we employ the efficient algorithms [Wong et al. 2017] to locate the means of random *PSD* matrices on the manifold and apply these concepts to the problem of *EEG* ischemic stroke detection. Thus the major contributions of this thesis can be summarized as follows:

- Feature extraction: we evaluate the cross *PSD* matrices of different *EEG* signal epochs in different sub-bands.
- Mean of *PSD* matrices: we obtain the means of the random *PSD* matrices of two different classes of *EEG* signals: $\mathcal{C}_0$ – those of healthy patients, and $\mathcal{C}_1$ – those of stroked patients.
- Optimum weighting: we use the collected patient *EEG* signals as training sets to obtain optimized weighting matrices for the RD.
- RD and weighted RD: we use the RD and the weighted RD to measure the distance and apply this concept to different classifiers to determine stroke condition.

Chapter 2

Signal Features for Classification

2.1 Pre-processing of *Electroencephalography* (*EEG*) signals

The measurement of electroencephalogram (EEG) is carried out by placing electrodes on different parts of the scalp, recording the electrical potentials generated by synaptic fields in the cerebral cortex. Although the electrodes would pick up the superposition of many different waves emitted from various regions of the brain, rendering the data more difficult to interpret, EEG is still a unique and valuable measure of the brain's electrical function.

For our purpose of detecting ischemic stroke, we collect from the patient EEG signal which is sampled at 256 Hz. The collected signal then goes through data preprocessing which consists of several steps: data referencing, segmentation, artifact removal, noise filtering, normalization and data collection.

2.1.1 Data Referencing

The EEG data are collected by placing electrodes at six locations, also called six channels: C3, C4, O1, O2, and two reference channels behind two ears as illustrated in (Fig.2.1).

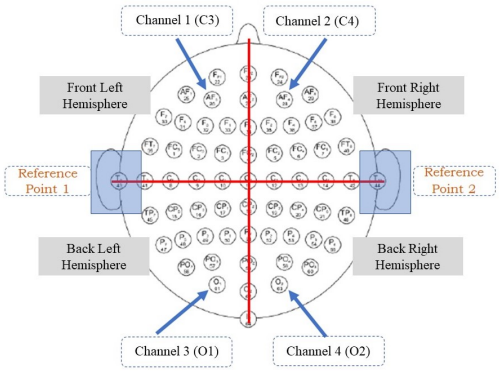

FIGURE 2.1: *EEG Channels with blue arrows are C3, C4, O1, O2 in front and back hemisphere.*

Mean of reference channels is first subtracted from data C3, C4, O1, O2,

$$x_c(n) = r_c(n) - \frac{r_{ref_1}(n) + r_{ref_2}(n)}{2} \tag{2.1}$$

where $x_c(n)$ representing the referenced EEG signals of channel $c \in \{C3, C4, O1, O2\}$, and $r_c(n)$ representing the raw EEG signals in channel $c \in \{C3, C4, O1, O2\}$, $r_{ref_1}(n)$ and $r_{ref_2}(n)$ refer to raw EEG signals in the two reference channels.

2.1.2 Segmentation

The EEG data in the 4 channels collected from a healthy participant or a stroke patient may last for hours in duration. In general, the signals are non-stationary. However, it is widely accepted that if we divide them into 30-second epochs, each epoch of the measured EEG data represents a wide-sense stationary signal. Hence, after data referencing, the signal in each channel is segmented into 30-second epochs having 7680 samples each.

2.1.3 Artifact Rejection

Artifact recognition and elimination is one of the most challenging part of monitoring *EEG*. Artifacts are caused by two factors: a) patient related artifacts (e.g. movement, sweating, ECG, eye movements) and b) technical artifacts (50/60 Hz artifact, cable movements, electrode paste-related) [mcgill 2018].

The presence of artifacts in *EEG* signals can distort the features which represent information of brain stroke and therefore leads to false detection results.

A common and effective clinical practice for artifact rejection for EEG signal is to identify artifacts by visual inference and then to manually remove the artifact signals. Other methods have been proposed to remove artifacts from EEG recordings including regression in time/frequency, and linear decomposition and reconstruction, etc. [Gratton et al. 1983, Woestenburg et al. 1983, Yoo et al. 2007, Shao et al. 2009, Anderson et al. 2006, Guerrero-Mosquera and Vazquez 2009].

There are existing tools for finding the artifacts. For example, FEMG and impedance measurements can be used for indicating contaminated signal. By looking at different parameters on a monitor, other interference may be found [mcgill 2018].

Since artifact rejection is not the main focus of this thesis, we adopt the visual inspection and manual artifact rejection method. The procedure is given in the following:

The magnitude of artifact usually is very large in a short duration compared to a normal brain wave signal. An artifact example shown in Fig.2.2 in which an artifact occurs during timestampe 100-200.

Therefore, for each epoch and for each single channel signal, we calculate the mean magnitude μ and standard derivation δ. Treating the distribution of the magnitude of *EEG* signal as Gaussian distribution, any sample which has magnitude larger than $|\mu + 3\delta|$ [Barnett and Lewis 1974], we remove it and replace by a random sample within the range of $\pm|\mu+3\delta|$. An example of *EEG* signal with artifact removed shown in Fig.2.3

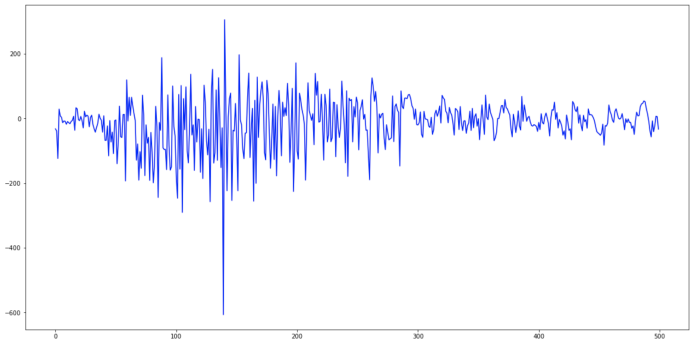

FIGURE 2.2: *EEG signal with artifact.*

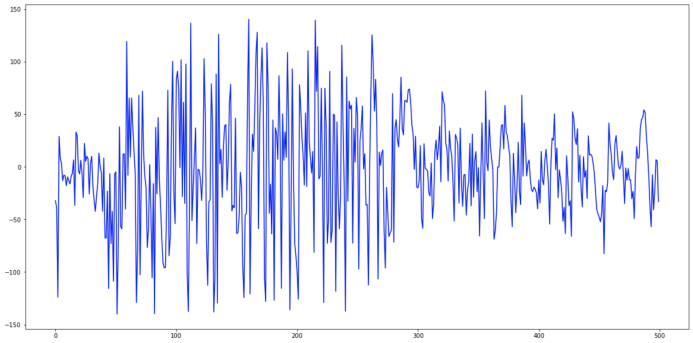

FIGURE 2.3: *EEG signal with artifact removed.*

2.1.4 Noise Filtering

Since for our application, all the *EEG* signals concentrate in the frequency range of $0-13$ Hz, therefore, we apply low pass filtering to the signal epochs after artifact have been removed to reduce the noise in the recorded *EEG* signals. The Butterworth filter design has a maximally flat amplitude response and relatively linear phase response in the pass-band [Bianchi and Sorrentino 2007]. Therefore, we choose a tenth order low-pass Butterworth filter with cut-off frequency of 58 Hz to ensure a relatively low distortion to signals.

2.1.5 *EEG* signal normalization and data collection

The EEG signal is recorded by electrodes placed at different locations of scalp. For the detection of ischemic brain stroke, $M = 4$ electrodes located at $C3$, $C4$, $O1$, $O2$ are used to collect multichannel time series signal.

Let the nth epoch of the mth channel measured signal be $\{s'_{nm}(t), m = 1, \cdots, M\}$ at the time instant t, we can represent the nth epoch of these multi-channel data at t as a vector: $\mathbf{s}'_n(t) = [s'_{n1}(t), \cdots, s'_{nM}(t)]^T,\ t = 1, \cdots, T$. Thus, the nth epoch measured data matrix (representing M channels of measured data for a duration of T seconds) for the patient is given by

$$\mathbf{S}'_n = [\mathbf{s}'_n(1), \cdots, \mathbf{s}'_n(T)], \qquad n = 1, \cdots, N \tag{2.2}$$

where $n = 1, ..., N$ and N is the number of number of epochs for a person.

The normalized EEG signal by using Frobenius norm is

$$\mathbf{S}_n = \frac{\mathbf{S}'_n}{||\mathbf{S}'_n||_F} = \frac{\mathbf{S}'_n}{(\sum_{i=1}^{M}\sum_{j=1}^{T}|[\mathbf{S}'_n]_{i,j}|^2)^{\frac{1}{2}}}. \tag{2.3}$$

For each patient, the labeled sample EEG signal is

$$\mathcal{L} = \left\{ \begin{bmatrix} \mathbf{S}_1 \\ l \end{bmatrix}, ..., \begin{bmatrix} \mathbf{S}_n \\ l \end{bmatrix}, ..., \begin{bmatrix} \mathbf{S}_N \\ l \end{bmatrix} \right\}, \tag{2.4}$$

where $l \in (0, 1)$. 0 represents that the epoch is belonged to a healthy person, and 1 represents that the epoch is belonged to a stroke patient.

2.2 Different Features Commonly Used for Classification

Here, we briefly present some of the commonly used signal features for stroke detection by *EEG*.

2.2.1 Frequency Domain Features

According to related medical literature, occurrences of ischemic strokes affect the signals in low frequency range (for example delta, theta and alpha bands) [Finnigan et al. 2016, Omar et al. 2014b]. An ischemic stroke usually occurs in a specific area (hemisphere) of the brain. Thus, there will be differences in relative power among different channels [Finnigan et al. 2016]. Therefore, changes in the power of each sub-band of channels and the relative power ratio between channels provide good indicators for the occurence of ischemic brain stroke and thus are suitable signal features for the process of detection.

Sub-band Power Estimation

For each epoch of single channel EEG signal $x(n)$, $n = 0, ..., N-1$, we used the Welch's weighted overlapped segment averaging method [Welch 1967] to estimate power spectral density. This involves splitting the recorded signal into overlapping windows of length L, calculating modified periodograms of these windows, and averaging these modified periodograms.

The resulting modified periodogram for the ith window is

$$\bar{p}^i(f) = \frac{1}{LU}\left|\sum_{n=0}^{L-1} x_i(n)w(n)e^{-j2\pi fn}\right|^2, \tag{2.5}$$

where U is the normalization factor for the power in the window function such that

$$U = \frac{1}{L}\sum_{n=0}^{L-1} w^2(n) \tag{2.6}$$

and $w(n)$ is the window function. The Welch power spectrum is the average of these modified periodograms:

$$\bar{p}(f) = \frac{1}{K}\sum_{i=0}^{K-1} \bar{p}^i(f) \tag{2.7}$$

In this paper, we used a 50% overlapped sliding window with 2 second window length, i.e., $L = 2 \times f_s = 512$ samples, to estimate the Power Spectral Density. The number of windows in an epoch is thus $T - 1 = 29$ and the average power of sub-band $b \in \{\delta, \alpha, \theta\}$ for channel $c \in \{C3, C4, O1, O2\}$ is given by:

$$P_b^c = \frac{\sum_{f_{min} \leq f < f_{max}} \bar{p}(f)}{n_b}, \tag{2.8}$$

where f_{min} and f_{max} are respectively the lower and upper of frequency range of each sub-band shown in (Fig. 1.2) , $n_b = (f_{max} - f_{min})/f_r$ is the number of frequency samples for each sub-band range. i.e. $f_r = f_s/L$.

Relative Band Power

From PSD, frequency bands and their relative strengths are known. Now for each channel, the relative power for each sub-band is given by:

$$\bar{P}_b^c = P_b^c \Big/ \sum_{c \in C} P_b^c, \tag{2.9}$$

where $c \in \{C3, C4, O1, O2\}$, $b \in \{\delta, \theta, \alpha\}$, and P_b^c is the average power of sub-band b recorded by channel c. Since we have three sub-bands and four channels for each observation, the total number of relative band power features is twelve.

Relative (Left and Right) Hemisphere Power

It shows the difference between the left hemisphere and the right hemisphere of person (fig. 2.1). In our EEG recording device, C3 and C4 channels located at the front left and the front right hemisphere respectively, O1 and O2 channels located at the back left and the back right hemisphere respectively. We calculate relative front hemisphere power $RPR(b)_{fh}$ the difference between C3 and C4 for each sub-band in (2.10):

$$(|\bar{P}_b^{C3} - \bar{P}_b^{C4}|) \Big/ (\bar{P}_b^{C3} + \bar{P}_b^{C4}), \tag{2.10}$$

where $\bar{P}_b^{C3}$ and $\bar{P}_b^{C4}$ are average power of sub-band $b \in \{\delta, \theta, \alpha\}$ in channel $C3$ and $C4$ respectively.

Similarly, we calculate relative back hemisphere power $RPR(b)_{bh}$ the difference between O1 and O2 for each sub band in (2.11):

$$(|\bar{P}_b^{O1} - \bar{P}_b^{O2}|)\Big/(\bar{P}_b^{O1} + \bar{P}_b^{O2}), \tag{2.11}$$

where $\bar{P}_b^{O1}$ and $\bar{P}_b^{O2}$ are average power of sub-band $b \in \{\delta, \theta, \alpha\}$ in channel $O1$ and $O2$ respectively.

Combining (2.9), (2.10) and (2.11), there are 18 possible frequency-domain features in each epoch that can be used for stroke detection. The frequency-domain features described above are the most commonly used EEG signal features used for detection of stroke.

2.2.2 Other Commonly Used Features

There are other suggestions such as:

- *Time-Domian Features* which mainly examine the *scale-invariant fluctuations* (showing specific brain activities) [Liu et al. 2016], thereby evaluating the *Hurst exponent* (h) which defines the particular kind of scale-invariant structure and fluctuation level in the EEG epochs of patient data; and
- *Time-Frequency Domain Features* which uses time-frequency domain spectrograms as features to train the Convolutional Neural Network (CNN) for mapping the changes in the EEG signal during the ischemic stroke detection [Matic et al. 2015]. The spectrogram is a visual representation of a relationship between frequency strength at specific time step. Spectrogram of each time window is used as an input for the training of Convolutional Neural Network.

2.3 *Power Spectral Density* (*PSD*) matrix for multi-channel signals

For a *Wide Sense Stationary* (*WSS*) stochastic signal $s(t)$, we have

$$\mathrm{E}[s(t)] = \mu = \text{constant} \tag{2.12}$$

$$\mathrm{E}[s(t+\tau)s(t)] = r_{ss}(\tau) \tag{2.13}$$

$r_{ss}(\tau)$ is called the autocorrelation of the signal $s(t)$. The power spectral density (also called the *power spectrum*) of $s(t)$ is the Fourier transform of the autocorrelation so that

$$p_{ss}(\omega) = \int_{-\infty}^{\infty} r_{ss}(\tau)e^{-j\omega\tau}d\tau \tag{2.14}$$

For two real WSS signals $s_1(t)$ and $s_2(t)$, the cross-correlation function is defined as

$$r_{s_1s_2}(\tau) = \mathrm{E}[s_1(t+\tau)s_2(t)] \tag{2.15}$$

We note that

$$r_{s_1s_2}(\tau) = r_{s_2s_1}(-\tau) \tag{2.16}$$

The cross-correlation and the cross-power spectral density are also related by the Fourier transform pair such that

$$p_{s_1s_2}(\omega) = \int_{-\infty}^{\infty} r_{s_1s_2}(\tau)e^{-j\omega\tau}d\tau \tag{2.17}$$

$$\tag{2.18}$$

Due to the anti-symmetry relationship of $r_{s_1s_2}$ and $r_{s_2s_1}(-\tau)$, their power spectra are conjugate pairs, i.e.,

$$p_{s_1s_2}(\omega) = p^*_{s_2s_1}(\omega) \tag{2.19}$$

Our work in this thesis employs a multi-channel EEG system such that for the nth epoch the EEG signals are in discrete-time forming a signal matrix $\mathbf{S}_n$ given by Eq. (2.3), where the column vector $\mathbf{s}_n(t) = [s_{n_1}(t), \ldots, s_{nM}(t)]^T$, $t = 1, \ldots, T$, is the measurement of the M channels at t, and can be considered as a *WSS* vector. Therefore, we can obtain its covariance matrix as follows:

We now vectorize this matrix so that $\tilde{\mathbf{s}} = \text{vec}(\mathbf{S}_n)$ is a $MT \times 1$ vector. The expected value of this vector can be found by the time average. This yields a vector consisting of T subvectors of M dimensions each:

$$\text{E}[\text{vec}(\mathbf{S}_n)] = \text{E}\begin{bmatrix} s_{1n}(0) \\ \vdots \\ s_{Mn}(0) \\ s_{1n}(1) \\ \vdots \\ s_{Mn}(1) \\ \vdots \\ s_{1n}(T-1) \\ \vdots \\ s_{Mn}(T-1) \end{bmatrix} \approx \tilde{\boldsymbol{\mu}}_n = \begin{bmatrix} \tilde{\mu}_{1n} \\ \vdots \\ \tilde{\mu}_{Mn} \\ \tilde{\mu}_{1n} \\ \vdots \\ \tilde{\mu}_{Mn} \\ \vdots \\ \tilde{\mu}_{1n} \\ \vdots \\ \tilde{\mu}_{Mn} \end{bmatrix} \tag{2.20}$$

$$\text{E}[s_{mn}(t)] \approx \tilde{\mu}_{mn} = \frac{1}{T}\sum_{t=0}^{T-1} s_{mn}(t) = \overline{s_{mn}(t)} \tag{2.21}$$

Its covariance matrix is $\tilde{\mathbf{K}}_n = \text{E}[(\tilde{\mathbf{s}}_n - \tilde{\boldsymbol{\mu}}_n)(\tilde{\mathbf{s}}_n - \tilde{\boldsymbol{\mu}}_n)^T] \approx \overline{[(\tilde{\mathbf{s}}_n - \tilde{\boldsymbol{\mu}}_n)(\tilde{\mathbf{s}}_n - \tilde{\boldsymbol{\mu}}_n)^T]}$ which is $MT \times MT$ and contains the $M \times M$ matrices $\mathbf{R}_n(\tau), \tau = 0, \cdots, T-1$, i.e.,

$$\tilde{\mathbf{K}}_n = \begin{bmatrix} \mathbf{K}_n(0) & \mathbf{K}_n(1) & \cdots & \mathbf{K}_n(T-1) \\ \mathbf{K}_n(-1) & \mathbf{K}_n(0) & \cdots & \mathbf{K}_n(T-2) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{K}_n(1-T) & \mathbf{K}_n(2-T) & \cdots & \mathbf{K}_n(0) \end{bmatrix}$$

where $\mathbf{K}_n(\tau) = [\kappa_{ij}(\tau)]$ with

$$\mathbf{K}_n(\tau) = \mathbf{K}_n(t_1 - t_2) = \begin{bmatrix} \kappa_{11}(\tau) & \kappa_{12}(\tau) & \cdots & \kappa_{1M}(\tau) \\ \kappa_{21}(\tau) & \kappa_{22}(\tau) & \cdots & \kappa_{2M}(\tau) \\ \vdots & \vdots & \vdots & \vdots \\ \kappa_{M1}(\tau) & \kappa_{M2}(\tau) & \cdots & \kappa_{MM}(\tau) \end{bmatrix}$$

where $\kappa_{ij}(\tau) = \mathrm{E}[(s_i(t) - \mu_i)(s_j(t+\tau) - \mu_j)]$. For real signals, $\kappa_{ij}(\tau) = \kappa_{ji}(-\tau)$. At any frequency ω, the power spectral density (PSD) matrix $\mathbf{P}_n(\omega)$, of the nth epoch signal is then the DFT

$$\mathbf{P}_n(\omega) = \sum_{\tau} e^{-j\omega\tau} \mathbf{K}_n(\tau) \tag{2.22}$$

Theoretically, the range of time-shift τ in the sum of 2.22 is $(-\infty, \infty)$. In practice, $\tau \in [-(T-1), T-1]$ since the number of samples is finite. Due to the anti-symmetry property of Eq. (2.19), $\kappa_{ij} = \kappa_{ji}(-\tau)$, and therefore, from Eq. (2.19), $\mathbf{P}_n(\omega)$ is a Hermitian matrix, i.e.,

$$\mathbf{P}_n(\omega) = \mathbf{P}_n^H(\omega) \tag{2.23}$$

It can be shown [Larsen 2015] that both $\mathbf{K}_n$ and $\mathbf{P}_n$ are positive semi-definite.

In our application, the PSD matrix of Eq. (2.22) is obtained using the Nuttall-Strand algorithm [Nuttall 1976, Strand 1977a], which is an accurate positive semi-definite estimation of the *PSD* matrix with high frequency resolution (see Section 2.5).

2.4 The *PSD* matrix and Its Manifold

The PSD matrix of Eq. (2.22) is, ingeneral, an $M \times M$ positive semi-definite Hermitian matrix. Some of its important properties which are used often in this thesis are listed below [Graybill 1983]:

1. The eigenvalues of an $M \times M$ Hermitian matrix $\mathbf{P} = \mathbf{P}^H$ are real and positive semi-definite. Furthermore, the eigenvectors belonging to the distinct eigenvalues

are orthogonal.

2. An $M \times M$ Hermitian matrix $\mathbf{P}$ can always be reduced to a diagonal matrix by unitary transformation, i.e., $\mathbf{U}^{-1}\mathbf{P}\mathbf{U} = \mathbf{\Lambda}$ where $\mathbf{U}^{-1} = \mathbf{U}^{H}$ and $\mathbf{\Lambda} = \text{diag}[\lambda_1, \ldots, \lambda_M]$.

3. For $\mathbf{P}$ being positive semi-definite Hermitan, there exists a unique positive semi-definite Hermitian matrix $\mathbf{P}^{1/2}$ such that $\mathbf{P}^{1/2}\mathbf{P}^{1/2} = \mathbf{P}$. $\mathbf{P}^{1/2}$ is called the square-root of $\mathbf{P}$.

The PSD Manifold

Consider the feature PSD matrices $\mathbf{P}$ in the signal space. These $M \times M$ matrices are Hermitian and positive definite, thus, they form a subset of $M \times M$ complex matrices. Suppose we denote the set of all the $M \times M$ complex matrices by $\mathcal{H}$. Also, we denote the set of all Hermitian matrices and the set of positive definite Hermitian matrices by $\mathcal{H}_{\text{H}}$ and $\mathcal{M}$, respectively. Thus, we have $\mathcal{M} \in \mathcal{H}_{\text{H}}$. Then, the following is an important property of $\mathcal{H}_{\text{H}}$ and $\mathcal{M}$:

Property 1. *$\mathcal{M}$ is a manifold[1] in the* real *linear vector space $\mathcal{H}_{\text{H}}$.* □

The proof of the above proposition is given in [Li and Wong 2013]

Property 1 is important in the development of distance measures in the manifold of complex PSD matrices. This is because we only have to consider real analysis of the geometry.

2.5 How are *PSD* matrices obtained from EEG signals

There are several ways to estimate *PSD* matrix for a multi-channel signal including Non-parametric methods and parametric methods. Non-parametric methods are simple, but are in general, not consistent in the estimation of the power spectrum. Parametric

[1] For now, a manifold can be looked upon as **Flanders** "An n-dimensional manifold is a space which is not necessarily a Euclidean space nor is it a domain in a Euclidean space, but which, from the viewpoint of a short-sighted observer living in the space, looks just like such a domain of Euclidean space."

methods are able to estimate *PSD* with higher accuracy if the model is chosen appropriately. There are many choices for non-parametric modeling like the Welch's method discussed in section 2.2.1, and parametric modeling including *Vector Auto-Regression* (*VAR*) model which will be discussed in the following.

2.5.1 *VAR* Model

In this thesis, we use the *VAR* model for the estimation of multi-channel *EEG* signal *PSD*, the outline shown in the following: The autocorrelation function (ACF) of a spectrally white multichannel noise sequence $\mathbf{n}(t)$ satisfies

$$\mathbf{R}_{nn}(\tau) = \mathtt{E}[\mathbf{n}^H(t)\mathbf{n}(t-\tau)] = \mathbf{P}_{nn}\delta(\tau) \tag{2.24}$$

where $\mathbf{P}_{nn}$ is a constant $M \times M$ matrix. Thus its *PSD* matrix is a constant, i.e.,

$$\mathbf{P}_{nn}(\omega) = \mathbf{P}_{nn}. \tag{2.25}$$

Now, the output signal of a q-th order *VAR* model can be described as

$$\mathbf{s}(t) = -\sum_{\tau=1}^{q} \mathbf{A}(\tau)\mathbf{s}(t-\tau) + \mathbf{n}(t) \tag{2.26}$$

where $\mathbf{A}(\tau)$ are the $M \times M$ coefficient matrices and $\mathbf{n}(t)$ is the $M \times 1$ vector of a spectrally white noise. Let $\mathbf{A}(0) = \mathbf{I}$. Then, the *Autocorrelation Function* (*ACF*) of $\mathbf{n}(t)$ is

$$\begin{aligned}
\mathbf{R}_{nn}(\tau) &= \mathtt{E}[\mathbf{n}^H(t)\mathbf{n}(t-\tau)] \\
&= \mathtt{E}[\sum_{k=0}^{q}\sum_{l=0}^{q} \mathbf{A}(k)\mathbf{s}(t-k)\mathbf{s}^T(t+k-l)\mathbf{A}^T(l)] \\
&= \sum_{k=0}^{q}\sum_{l=0}^{q} \mathbf{A}(k)\mathbf{R}_{ss}(\tau+k-l)\mathbf{A}^T(l)
\end{aligned} \tag{2.27}$$

Taking the z-transform of (2.27), we have

$$\begin{aligned}\mathcal{Z}[\mathbf{R}_{nn}(\tau)] &= \sum_{\tau=-\infty}^{\infty} \mathbf{R}_{nn}(\tau) z^{-\tau} \\ &= (\sum_{k=0}^{q} \mathbf{A}(k) z^{k})(\sum_{\tau=-\infty^{\infty}} \mathbf{R}_{ss}(\tau+k-l) z^{\tau+k-l} z^{-}(\tau+k-l))(\sum_{l=0}^{q} \mathbf{A}^{T}(l) z^{-l}).\end{aligned} \tag{2.28}$$

Let $z = e^{j\omega}$ and use (2.22), we have

$$\mathbf{P}_{nn}(\omega) = (\sum_{k=0}^{q} \mathbf{A}(k) e^{j\omega k}) \mathbf{P}_{ss}(\omega) (\sum_{l=0}^{q} \mathbf{A}^{T}(l) e^{-j\omega l}). \tag{2.29}$$

Therefore, we have

$$\begin{aligned}\mathbf{P}_{ss}(\omega) &= (\sum_{k=0}^{q} \mathbf{A}(k) e^{j\omega k})^{-1} \mathbf{P}_{nn}(\omega) (\sum_{l=0}^{q} \mathbf{A}^{T}(l) e^{-j\omega l})^{-1} \\ &= (\sum_{k=0}^{q} \mathbf{A}(k) e^{j\omega k})^{-1} \mathbf{P}_{nn} (\sum_{l=0}^{q} \mathbf{A}^{T}(l) e^{-j\omega l})^{-1}\end{aligned} \tag{2.30}$$

by (2.25). Let

$$\mathbf{A}(\omega) = \sum_{\tau=0}^{q} \mathbf{A}(\tau) e^{-j\omega\tau} \tag{2.31}$$

Then

$$\begin{aligned}\mathbf{A}^{H}(\omega) &= (\sum_{\tau=0}^{q} \mathbf{A}(\tau) e^{-j\omega\tau})^{H} \\ &= \sum_{\tau=0}^{q} \mathbf{A}^{H}(\tau) (e^{-j\omega\tau})^{H} \\ &= \sum_{\tau=0}^{q} \mathbf{A}^{T}(\tau) e^{j\omega\tau}\end{aligned} \tag{2.32}$$

Thus, the (2.30) can be rewritten as

$$\mathbf{P}_{ss}(\omega) = \mathbf{A}^{-1}(-\omega) \mathbf{P}_{nn} \mathbf{A}^{-H}(\omega). \tag{2.33}$$

From (2.33) we see that to find the power spectral density matrices $\mathbf{P}_{ss}(\omega)$ of the signal $\mathbf{s}(t)$ one needs to estimate the coefficient matrices $\mathbf{A}(\tau)$ in the VAR model of (2.26). We can use Nuttall-Strand algorithm, which is well-known to estimate the coefficient matrices $\mathbf{A}(\omega)$ and PSD $\mathbf{P}_{nn}$ of the spectrally white noise by observed signal sequence.

2.5.2 The Nuttall-Strand Algorithm

We consider the forward and backward filters which are multichannel AR models [Haykin 2008] of order q, i.e.,

$$\mathbf{e}_q(t) = \mathbf{s}(t) + \sum_{k=1}^{q} \mathbf{A}(k)\mathbf{s}(t-k) \tag{2.34}$$

and

$$\mathbf{b}_q(t) = \mathbf{s}(t) + \sum_{k=1}^{q} \mathbf{B}(k)\mathbf{s}(t+k) \tag{2.35}$$

respectively. The optimum forward and backward filters can be obtained by minimizing the expected mean-square values of $\mathbf{e}_q(t)$ and $\mathbf{b}_q(t)$. The minimum of $\mathtt{E}[\mathbf{e}_q^T(t)\mathbf{e}_q(t)]$ leads to the equations:

$$\mathbf{R}^{fw}\mathbf{F}(q) = \mathbf{V}^{fw} \tag{2.36}$$

where

$$\mathbf{F}(q) = [\mathbf{I}, \mathbf{A}^T(1), \ldots, \mathbf{A}^T(q)]^T \tag{2.37}$$

$\mathbf{R}^{fw} = [\mathbf{R}_{ik}^{fw}]$, where $\mathbf{R}_{ik}^{fw} = \mathbf{R}_{k-i}^{fw}$, $i,k = 1,2,...,q$, and $\mathbf{V}^{fw} = [\mathbf{P}^{fw}, \mathbf{0}, \ldots, \mathbf{0}]^T$ with $\mathbf{P}^{fw} = \mathtt{E}[\mathbf{e}_q^T(t)\mathbf{e}_q(t)]$. Similarly, the minimum of $\mathtt{E}[\mathbf{b}_q^T(t)\mathbf{b}_q(t)]$ leads to the equation:

$$\mathbf{R}^{bw}\mathbf{B}(q) = \mathbf{V}^{bw} \tag{2.38}$$

where

$$\mathbf{B}(q) = [\mathbf{I}, \mathbf{B}^T(1), \ldots, \mathbf{B}^T(q)]^T \tag{2.39}$$

$\mathbf{R}^{bw} = [\mathbf{R}_{ik}^{bw}]$, where $\mathbf{R}_{ik}^{bw} = \mathbf{R}_{k-i}^{bw}$, $i,k = 1,2,...,q$, and $\mathbf{V}^{bw} = [\mathbf{P}^{bw}, \mathbf{0}, \ldots, \mathbf{0}]^T$ with $\mathbf{P}^{bw} = \mathtt{E}[\mathbf{e}_q^T(t)\mathbf{e}_q(t)]$. To solve (2.36) and (2.38), the forward and backward filters may

be postulated as

$$\mathbf{F}(q) = \begin{bmatrix} \mathbf{F}(q-1) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{B}^{bw}(q-1) \end{bmatrix} \mathbf{C}^{fw}(q) \tag{2.40}$$

and

$$\mathbf{B}(q) = \begin{bmatrix} \mathbf{B}(q-1) \\ 0 \end{bmatrix} \mathbf{C}^{bw}(q) + \begin{bmatrix} 0 \\ \mathbf{B}^{bw}(q-1) \end{bmatrix} \tag{2.41}$$

where $\mathbf{B}^{bw}(q-1) = [\mathbf{B}^T(q-1), \ldots, \mathbf{B}^{bw}(1), \mathbf{I}]^T$.

Let $\{\mathbf{s}_t : t = 1, \ldots, T\}$ be a sample of T consecutive observations of the *EEG* signal. Let

$$\mathbf{s}_k(q) = [\mathbf{s}_{k+q}^T, \mathbf{s}_{k+q-1}^T, \ldots, \mathbf{s}_k^T]^T \tag{2.42}$$

for $k = 1, 2, \ldots, N-q$, $q = 0, 1, \ldots, T-1$. Let

$$\mathbf{e}_k(q) = [(\mathbf{F}(q-1))^T, \mathbf{0}]\mathbf{s}_k(q) \tag{2.43}$$

and

$$\mathbf{b}_k(q) = [\mathbf{0}, (\mathbf{B}^{bw}(q-1))^T]\mathbf{s}_k(q). \tag{2.44}$$

Then, the algorithm is as follows:

Algorithm (Nuttall-Strand)

1. Initialize the residual power matrices $\mathbf{P}^{fw}(0)$ and $\mathbf{P}^{bw}(0)$:

$$\mathbf{P}^{fw}(0) = \mathbf{P}^{bw}(0) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{s}_t \mathbf{s}_t^T \tag{2.45}$$

2. Calculate the forward and backward residuals for $k = 1, \ldots, T-q$:

   (a) $q = 1$: $\mathbf{e}_k(q) = \mathbf{s}_{k+1}$, $\mathbf{b}_k(q) = \mathbf{s}_k$

   (b) $q > 1$: $\mathbf{e}_k(q) = \mathbf{e}_{k+1}(q-1) + (\mathbf{C}^{fw}(q-1))^T \mathbf{b}_{k+1}(q-1)$, $\mathbf{b}_k(q) = \mathbf{b}_k(q-1) + (\mathbf{C}^{bw}(q-1))^T \mathbf{e}_k(q-1)$

3. Calculate

$$\mathbf{E} = \frac{1}{T-q}\sum_{k=1}^{T-q}\mathbf{e}_k(q)\mathbf{e}_k^T(q) \tag{2.46}$$

$$\mathbf{G} = \frac{1}{T-q}\sum_{k=1}^{T-q}\mathbf{b}_k(q)\mathbf{e}_k^T(q) \tag{2.47}$$

$$\mathbf{B} = \frac{1}{T-q}\sum_{k=1}^{T-q}\mathbf{e}_k(q)\mathbf{b}_k^T(q) \tag{2.48}$$

4. Solve $\mathbf{c}_{fw}(q)$ from

$$\mathbf{B}\mathbf{C}^{fw}(q) + \mathbf{P}^{bw}(q-1)\mathbf{C}^{fw}(q)(\mathbf{P}^{fw}(q-1))^{-1}\mathbf{E} = -2\mathbf{G} \tag{2.49}$$

5. Compute $\mathbf{c}_{bw}(q)$ by

$$\mathbf{c}_{bw}(q) = (\mathbf{P}^{fw}(q-1))^{-1}\mathbf{C}^T(q)\mathbf{P}^{fw}(q-1) \tag{2.50}$$

6. Computer power matrices $\mathbf{P}^{fw}(q)$ and $\mathbf{P}^{bw}(q)$ by

$$\mathbf{P}^{fw}(q) = \mathbf{P}^{fw}(q-1) - (\mathbf{C}^{fw}(q))^T\mathbf{P}^{fw}(q-1)\mathbf{C}^{fw}(1) \tag{2.51}$$

and

$$\mathbf{P}^{bw}(q) = \mathbf{P}^{bw}(q-1) - (\mathbf{C}^{bw}(q))^T\mathbf{P}^{fw}(q-1)\mathbf{C}^{bw}(1) \tag{2.52}$$

7. Update the filters coefficients using (2.40) and (2.41).

8. If $||\mathbf{P}^{fw}(q) - \mathbf{P}^{bw}(q)|| < \epsilon$, then go to 2.

9. Calculate the power spectral density matrix

$$\mathbf{P}(\omega) = \mathbf{A}^{-1}(-\omega)\mathbf{P}^{fw}(q)\mathbf{A}^{-T}(\omega) \tag{2.53}$$

where $\mathbf{A}(\omega) = \mathbf{I} + \mathbf{A}(1)e^{-j\omega} + \cdots + \mathbf{A}(q)e^{-j\omega}$.

For detailed derivation of the algorithm and the implementation, see [Strand 1977b, Nuttall 1976].

Chapter 3

The concept of distance in signal classification

In Chapter 2, we discussed how the pre-processing of *Electroencephalography* (*EEG*) signals can be carried out, including the cleaning up of artifacts and noise. We also discussed the process of feature extraction, and in particular, we discussed how the feature of the *Power Spectral Density* (*PSD*) matrix can be obtained. The major properties of these *PSD* matrices including how they form a manifold in the signal space are also reviewed. We will use these features for the purpose of signal classification (brain stroke detection) in the ensuing chapters.

In this chapter, we discuss the concept of similarity/dissimilarity between features of signals. The requirement of the measurement of similarity/dissimilarity is that the quantified similarity between features of signals should be small if the signals come from the same class and be large if the signals come from different classes.

Since *PSD* matrices can be represented as points in a linear signal space, the similarity between these features can then be intuitively measured by some kind of distance function, or distance for short. In this chapter, we will discuss different metrics to measure distance between points in these spaces: First, the definition of distance; second, different types of distance; third, how to establish distance between two *PSD* matrices.

3.1 Definition Of Distances

A general approach for characterizing the difference between two signals is to assign a "distance", a positive, real number, to each pair of elements of a signal. The function to evaluate distance between two elements of a signal set is called a *metric* which satisfies the following three properties:

$$d(x,y) \geq 0 \quad \text{and} \quad d(x,y) = 0 \quad \text{iff} \quad x = y \qquad (positivity) \tag{3.1}$$

$$d(x,y) = d(y,x) \qquad (symmetry) \tag{3.2}$$

$$d(x,z) \leq d(x,y) + d(y,z) \qquad (triangle\ inequality) \tag{3.3}$$

3.2 Distances Induced By Inner Product

In signal processing, the Euclidean (inner product) distance (ED) is the most commonly used distance measure [Franks 1969, Papoulis 1977] because it coincides with the usual concept of distance in a 3-dimensional space and also represents many important physical quantities. For two n-dimensional complex vectors $\mathbf{x}$ and $\mathbf{y}$, the ED is defined as [Franks 1969]

$$d_{\mathrm{E}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2} \tag{3.4}$$

From the Cauchy-Schwarz inequality, we have $|\langle \mathbf{x},\mathbf{y}\rangle| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$, so that a real angle θ between $\mathbf{x}$ and $\mathbf{y}$ can be defined as $\cos\theta = \frac{|\langle \mathbf{x},\mathbf{y}\rangle|}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$. Thus, another distance measure between two n-dimensional normalized vectors can be established based on their *correlation* such that the smaller the angle, the shorter is the distance[1], i.e.,

$$d_{\mathrm{C}}(\mathbf{x},\mathbf{y}) = \sqrt{1 - \frac{|\sum_{i=1}^{n} x_i y_i^*|}{\|\mathbf{x}\|\|\mathbf{y}\|}} \tag{3.5}$$

[1]The angle, $\theta = \cos^{-1}\left((|\langle \mathbf{x},\mathbf{y}\rangle|)/(\|\mathbf{x}\|\,\|\mathbf{y}\|)\right)$, can also be used as distance measures. If the normalized inner product is replaced by the product of two probability distributions (real and normalized), then θ defines the Fisher-Rao distance [Bengtsson and Życzkowski 2017]. If $\mathbf{x},\mathbf{y}$ are complex quantities whose modulus are probability distributions, 2θ defines the Fubini-Study distance [Kendall et al. 1946].

We may also look upon an $M \times M$ complex matrix as a point in the M^2 complex signal space so that the same idea of distance [Golub and Van Loan 2012] between two such matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ can also be applied:

$$d_{\mathrm{F}}(\mathbf{A},\mathbf{B}) = \left(\sum_{i=1}^{M} \sum_{j=1}^{M} |a_{ij} - b_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}[(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^H]} \tag{3.6}$$

Eq. (3.6) is often called the *Frobenius distance* which is in fact induced by the inner product norm and can be considered as the ED between $\mathbf{A}$ and $\mathbf{B}$ since, if we form the vectors vec$\mathbf{A}$ and vec$\mathbf{B}$ using the vec-functions [Horn et al. 1990], it is easy to see that $d_{\mathrm{F}}^2(\mathbf{A},\mathbf{B}) = d_{\mathrm{E}}^2(\mathrm{vec}\mathbf{A}, \mathrm{vec}\mathbf{B})$.

3.3 Distance Between Two PSD Matrices

3.3.1 Euclidean distance

We can apply Eq. (3.6) to two M by M complex PSD matrices $\mathbf{P}_m, \mathbf{P}_n$ giving

$$\begin{aligned} d_E(\mathbf{P}_m, \mathbf{P}_n) &= \left(\sum_{i=1}^{M} \sum_{j=1}^{M} |p_{m_{i,j}} - p_{n_{i,j}}|^2 \right)^{1/2} \\ &= \sqrt{tr[(\mathbf{P}_m - \mathbf{P}_n)(\mathbf{P}_m - \mathbf{P}_n)^H]}, \end{aligned} \tag{3.7}$$

where $p_{m_{i,j}}$ and $p_{n_{i,j}}$ are the ij-th wlwmwnts of $\mathbf{P}_m$ and $\mathbf{P}_n$ respectively. Eq. (3.7) gives the Euclidean between two PSD matrices $\mathbf{P}_m, \mathbf{P}_n$. We note that the Euclidean distance between two PSD matrices measeures the straight line distance between the two points in the signal space.

3.3.2 Riemannian distance for measurement of distance between two PSD matrices

Since the PSD matrices form a manifold in the signal space, therefore, measuring the distance between two PSD matrices should be carried out along the surface of the manifold. Thus, the Euclidean distance is not an appropriate measure for the similarity/dissimiliarity between two PSD matrices.

To find the distance between two points on the surface of the manifold, we have to employ knowledge from differential geometry. We gather that the length of a path between the two points on the manifold $\mathcal{M}$ is given by [Jost and Jost 2008]:

$$\ell(\mathbf{P}) = \int_{\theta_m}^{\theta_n} g_{\mathbf{P}}^{1/2}(\dot{\mathbf{P}}, \dot{\mathbf{P}})\ d\theta \tag{3.8}$$

where θ parameterizes the path joining $\mathbf{P}_m$ and $\mathbf{P}_n$, θ_m and θ_n being the values of the parameter at $\mathbf{P}_m$ and $\mathbf{P}_n$ respectively, $\dot{\mathbf{P}} = \frac{d\mathbf{P}}{d\theta}$, and $g_{\mathbf{P}}(\dot{\mathbf{P}}, \dot{\mathbf{P}})$ is an inner product metric, called a *Riemannian metric*, at $\mathbf{P}$ on $\mathcal{M}$, which can be defined in a variety of ways. A differentiable manifold $\mathcal{M}$ in which each tangent space is endowed with a Riemannian metric is called a *Riemannian manifold*. The curve on the manifold linking two PSD matrices $\mathbf{P}_m$ and $\mathbf{P}_n$ having the minimum length is called a *geodesic*, and the length of the geodesic is called the *Riemannian distance* (RD) between the two points, i.e.,

$$d_{\mathrm{R}}(\mathbf{P}_m, \mathbf{P}_n) \triangleq \min_{\mathbf{P}(\theta):[\theta_m,\theta_n]\to\mathcal{M}} \{\ell(\mathbf{P}(\theta))\} \tag{3.9}$$

Differently defined Riemannian metrics give rise to different RD. The direct evaluation of the RD in Eq. (3.9) is difficult. The following concept developed in [Li and Wong 2013] helps to solve the problem:

Let $\mathcal{H}$ denotes the Euclidean space of all $M \times M$ complex matrices. We can establish a mapping $\pi : \mathcal{M} \to \mathcal{H}$ associating each point $\mathbf{P} \in \mathcal{M}$ with $\tilde{\mathbf{P}} \triangleq \pi(\mathbf{P})$. Then, $\tilde{\mathbf{P}}$ is still an $M \times M$ complex matrix but may no longer be positive semi-definite or Hermitian, i.e., $\tilde{\mathbf{P}} \in \mathcal{H}$.

By choosing a particular mapping π, together with an appropriate Riemannian metric, we can find a Euclidean subspace $\mathcal{U}_{\mathcal{H}}$ at $\tilde{\mathbf{P}}$ of $\mathcal{H}$ which is *isometric* with $\mathcal{T}_{\mathcal{M}}(\mathbf{P})$, the tangent space at $\mathbf{P}$ on the manifold $\mathcal{M}$, i.e., the geodesic between $\mathbf{P}_m, \mathbf{P}_n \in \mathcal{M}$ can be *lifted* to $\tilde{\mathbf{P}}_m, \tilde{\mathbf{P}}_n \in \mathcal{U}_{\mathcal{H}}$. Thus, the RD on the manifold can be expressed directly in the Euclidean subspace $\mathcal{U}_{\mathcal{H}}$ in which ED is the measure distance. Three different closed-form expressions of RD for the PSD matrix manifold have been obtained following this method [Li and Wong 2013].

1. RD d_{R_1}: We use the mapping π such that $\mathbf{P} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^H$, i.e., $\tilde{\mathbf{P}} = \mathbf{P}^{1/2}\mathbf{U}$ where $\tilde{\mathbf{P}} \in \mathcal{H}$, $\mathbf{P} \in \mathcal{M}$, $\mathbf{U}$ is a unitary matrix, and choose the Riemannian metric on $\mathcal{M}$ as $g_{\mathbf{P}}(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\mathrm{tr}\mathbf{A}\mathbf{K}$ with $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$, and $\mathbf{K}\mathbf{P} + \mathbf{P}\mathbf{K} = \mathbf{B}$. Then, $\mathbf{P}_m$ and $\mathbf{P}_n$ can be lifted to $\tilde{\mathbf{P}}_m, \tilde{\mathbf{P}}_n \in \mathcal{U}_{\mathcal{H}}$ by letting $\tilde{\mathbf{P}}_m = \mathbf{P}_m^{1/2}\mathbf{U}_{rm}$ and $\tilde{\mathbf{P}}_n = \mathbf{P}_n^{1/2}\mathbf{U}_{\ell m}$, such that $\mathbf{P}_n^{1/2}\mathbf{P}_m^{1/2} = \mathbf{U}_{\ell m}\mathbf{\Sigma}\mathbf{U}_{rm}^H$ with $\mathbf{\Sigma}$ being the singular value matrix, and $\mathbf{U}_{\ell m}$ and $\mathbf{U}_{rm}$ being the left and right singular vector matrices of $\mathbf{P}_n^{1/2}\mathbf{P}_m^{1/2}$ respectively. The RD can be found to be

$$d_{\mathrm{R}_1}(\mathbf{P}_m, \mathbf{P}_n) = \sqrt{\mathrm{tr}\mathbf{P}_m + \mathrm{tr}\mathbf{P}_n - 2\mathrm{tr}\left[(\mathbf{P}_m^{1/2}\mathbf{P}_n\mathbf{P}_m^{1/2})^{1/2}\right]} \tag{3.10}$$

2. RD d_{R_2}: Use the mapping π such that $\mathbf{P} = \tilde{\mathbf{P}}^2$, and choose the Riemannian metric $g_P(\mathbf{A}, \mathbf{B}) = \langle \mathbf{A}, \mathbf{K} \rangle$ where $\mathbf{P}\mathbf{K} + \mathbf{K}\mathbf{P} + 2\tilde{\mathbf{P}}\mathbf{K}\tilde{\mathbf{P}} = \mathbf{B}$, with $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$, then the RD between $\mathbf{P}_m$ and $\mathbf{P}_n$ on $\mathcal{M}$ is:

$$d_{\mathrm{R}_2}(\mathbf{P}_m, \mathbf{P}_n) = \sqrt{\mathrm{tr}\mathbf{P}_m + \mathrm{tr}\mathbf{P}_n - 2\mathrm{tr}[\mathbf{P}_m^{1/2}\mathbf{P}_n^{1/2}]} \tag{3.11}$$

3. RD d_{R_3}: Use the mapping π: $\mathbf{P} = \exp(\tilde{\mathbf{P}})$, and choose the Riemannian metric $g_P(\mathbf{A}, \mathbf{B}) = \mathrm{tr}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}^{-1}\mathbf{B})$ where $\mathbf{A}, \mathbf{B} \in \mathcal{T}_{\mathcal{M}}(\mathbf{P})$, then the RD between $\mathbf{P}_m$ and $\mathbf{P}_n$ in $\mathcal{M}$ is

$$d_{\mathrm{R}_3}(\mathbf{P}_m, \mathbf{P}_n) = \sqrt{\mathrm{tr}[(\log \mathbf{P}_m^{-1/2}\mathbf{P}_n\mathbf{P}_m^{-1/2})^2]} = \sqrt{\sum_{i=1}^{M} \log^2 \lambda_i} \tag{3.12}$$

where λ_i are the eigenvalues of $\mathbf{P}_m^{-1}\mathbf{P}_n$.

Remarks:

a) All three RD satisfy the axioms of distance, i.e., (i) positivity, (ii) symmetry, (iii) triangle inequality.

b) RD d_{R_1} and d_{R_2} are expressions newly developed [Li and Wong 2013]. On the other hand, d_{R_3} which can be arrived at in several different ways, has been in use for a long time in physics and mathematics (especially in General Relativity Theory) [Bhatia 2009, Besse 2007, Moakher 2005].

c) In signal processing, $d_{\mathrm{R}3}$ has been studied for statistical operations and applied to interpolation, filtering, and restoration of PSD matrices [Arsigny et al. 2006, Arsigny et al. 2006]. As well, based on d_{R_3}, different classification algorithms have been developed and have been applied to the detection of pedestrians, MRI and EEG classifications [Tuzel et al. 2008, Barachant et al. 2012]. However, d_{R_3} is not sensitive to weighting and thus may not be appropriate for use when *a priori* information is available (see Section 3.3.3. On the other hand, d_{R_1} and d_{R_2} are not quite widely used yet, however, due to it being more mathematically manipulatable, d_{R_2} has recently been applied in robust beamforming and signal detection with rather attractive results [Ciochina et al. 2013, Xu et al. 2013, Wong et al. 2017]. Here in this thesis, we also employ the RD d_{R_2} for classification of stroke EEG signals.

3.3.3 Weighting of Riemannian Distances

Applying weighting matrix is a general way to enhance the similarity and dissimilarity between the features of *PSD* matrices. In order to do so, a positive definite Hermitian weighting matrix $\mathbf{W}$ can be applied to the *PSD* feature matrices such that $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H$, where $\mathbf{\Omega}$ is an $M \times K$, $K \leq M$ matrix. Then the weighted version of $\mathbf{P}_m$ and $\mathbf{P}_n$ can be defined as $\mathbf{P}_{m\mathrm{W}} = \mathbf{\Omega}^H\mathbf{P}_m\mathbf{\Omega}$ and $\mathbf{P}_{n\mathrm{W}} = \mathbf{\Omega}^H\mathbf{P}_n\mathbf{\Omega}$, respectively. It is easy to see that $\mathbf{P}_{m\mathrm{W}}$ and $\mathbf{P}_{n\mathrm{W}}$ are also positive semi-definite Hermitian matrices on the manifold. The distance between two weighted *PSD* matrices then results in a weighted

Riemannian Distance (*RD*). For the three *RD* d_{R_1}, d_{R_2}, and d_{R_3}, their corresponding weighted distances between $\mathbf{P}_m$ and $\mathbf{P}_n \in \mathcal{M}$ are respectively given by [Li and Wong 2013]

$$d_{\mathrm{WR}_1}(\mathbf{P}_m, \mathbf{P}_n) = \sqrt{\mathrm{tr}\mathbf{W}\mathbf{P}_m + \mathrm{tr}\mathbf{W}\mathbf{P}_n - 2\mathrm{tr}(\mathbf{P}_m^{1/2}\mathbf{W}\mathbf{P}_n\mathbf{W}\mathbf{P}_m^{1/2})^{1/2}} \tag{3.13}$$

$$d_{\mathrm{WR}_2}(\mathbf{P}_m, \mathbf{P}_n) = \sqrt{\mathrm{tr}\mathbf{W}\mathbf{P}_m + \mathrm{tr}\mathbf{W}\mathbf{P}_n - \mathrm{tr}\mathbf{W}\mathbf{P}_m^{1/2}\mathbf{P}_n^{1/2} - \mathrm{tr}\mathbf{W}\mathbf{P}_n^{1/2}\mathbf{P}_m^{1/2}} \tag{3.14}$$

$$d_{\mathrm{WR}_3}(\mathbf{P}_m, \mathbf{P}_n) = d_{R_3}(\mathbf{P}_{m\mathrm{W}}, \mathbf{P}_{n\mathrm{W}}). \tag{3.15}$$

We note from the above equation, that d_{R_3} is *weight-invariant*, meaning that the distance measure does not change with weighting. Thus, the use of d_{R_3} cannot benefit from the *a priori* knowledge and improve the similarity/dissimilarity in the process of classification. Thus, in this thesis, we will not include the study of using d_{R_3}.

3.4 Dissimilarity Measures of Two Epochs of EEG Signals

Now we are ready to define the similarity/dissimilarity between the *PSD* matrices of two epochs of EEG signals:

PSD is a function of the frequency ω. With the variation of ω, the *PSD* describes a curve of on the Riemannian manifold $\mathcal{M}$. Therefore, similarity/dissimilarity between two sequences of *PSD* matrices corresponding to two multi-channel epochs of *EEG* signals must be established.

For two curves on the manifold described by two *PSD* matrices $\mathbf{P}_m(\omega)$ and $\mathbf{P}_n(\omega)$, the distance, $d(\mathbf{P}_m(\omega), \mathbf{P}_n(\omega))$, ED or RD alike, is a non-negative real valued function of ω, measuring the distance between the two curves at frequency ω. As ω varies, we define the average distance between the curves $\mathbf{P}_m(\omega)$ and $\mathbf{P}_n(\omega)$ in the range of $[\omega_{\min}, \omega_{\max}]$

as

$$\begin{aligned}\bar{d}(\mathbf{P}_m(\omega),\mathbf{P}_n(\omega)) &= \frac{1}{(\omega_{\max}-\omega_{\min})}\int_{\omega_{\min}}^{\omega_{\max}} d(\mathbf{P}_m(\omega),\mathbf{P}_n(\omega))d\omega \\ &\approx \frac{1}{N}\sum_{i=1}^{N} d(\mathbf{P}_m(\omega_i),\mathbf{P}_n(\omega_i))\Delta\omega_i. \end{aligned} \tag{3.16}$$

where $d(\mathbf{P}_m(\omega_i),\mathbf{P}_n(\omega_i))$ can stand for either ED or RD. It is easy to show this Riemann integral satisfies the axioms of a distance function. If equal frequency increment is used, i.e., Δ_{ω_i} is a constant, then we can define the overall Euclidean and Riemannian dissimilarity between the two given *PSD* curves as

$$\hat{d}_{\mathrm{E}}(\mathbf{P}_m(\omega),\mathbf{P}_n(\omega)) = \textstyle\sum_i d_{\mathrm{E}}(\mathbf{P}_m(\omega_i),\mathbf{P}_n(\omega_i)) \tag{3.17a}$$

$$\hat{d}_{\mathrm{R}}(\mathbf{P}_m(\omega),\mathbf{P}_n(\omega)) = \textstyle\sum_i d_{\mathrm{R}}(\mathbf{P}_m(\omega_i),\mathbf{P}_n(\omega_i)) \tag{3.17b}$$

Therefore, we can evaluate the similarity/dissimilarity by (3.17) in the observed frequency range for the purpose of classification in either the Euclidean sense or the Riemannian sense.

3.5 Mean of Normalized Random PSD Matrices

The *mean* is a fundamental statistic used in signal processing to represent centrality of data points. For real scalars, the mean minimizes the sum of the squared distances from the points to this central point [Kendall et al. 1946].

We can generalize these properties to define the mean of $M \times M$ PSD matrices, $\{\mathbf{P}_n, n = 1, \cdots, N\}$, by using the geometric distance d measured between two matrices. Thus, we have,

$$\mathbf{C} = \arg\min_{\mathbf{C}} \sum_{n=1}^{N} d^2(\mathbf{C},\mathbf{P}_n) \tag{3.18}$$

where d, the distance measured between two matrices, is a general metric. If d represents the various RD d_{R}, then the central points are respectively called the *Riemannian mean* (RMn), denoted by $\mathbf{C}_{\mathrm{R}}$. Likewise, for d being ED, the central points are *Euclidean mean* (EMn), denoted by $\mathbf{C}_{\mathrm{E}}$.

3.5.1 Euclidean mean

For Eiclidean mean of a group of N random normalized PSD matrices $\{\mathbf{P}_n\}; n = 1, \cdots, N$, the mean is simply the arithmatic average [Kendall et al. 1946] such that

$$\mathbf{C}_{\mathrm{E}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{P}_n \tag{3.19}$$

That $\mathbf{C}_{\mathrm{E}}$ indeed minimizes the sum squared distance from all points can be shown by simply differentiating the the sum squared distances from all the points to $\mathbf{C}$ with respect to $\mathbf{C}$, and equate the result to zero for minimum. Eq. (3.19) then follows.

3.5.2 Riemannian mean

The RMn of PSD matrices have been studied by various researchers recently [Ning et al. 2013, Arnaudon et al. 2013, Moakher 2005, Wong et al. 2017]. In this section, we gives an exposition to the algorithm developed by [Wong et al. 2017] based on the concept of alternating mapping between base and total-spaces to find the RMn according to the specific distance measure of d_{R_2} for a group of identically distributed PSD matrices lying in a convex set $\mathcal{C} \subset \mathcal{M}$, i.e., between any two point in $\mathcal{C}$, there is a unique geodesic lying entirely in $\mathcal{C}$.

The algorithm for finding the RMn of a group of PSD matrices $\{\mathbf{P}_n\}; n = 1, \cdots, N$ is based on the concept that each pair of the points $\mathbf{P}_m$ and $\mathbf{P}_n$ can be systematically lifted to a Euclidean subspace $\mathcal{U}_{\mathcal{H}}$ isometric to the tangent space of the manifold $\mathcal{M}$, yielding the points $\tilde{\mathbf{P}}_m$ and $\tilde{\mathbf{P}}_n$. Then adjusting the Euclidean subspaces by finding the EMn of these lifted points $\{\tilde{\mathbf{P}}_n\}; n = 1, \cdots, N$ results in a $\tilde{\mathbf{C}}$ whose image on the

manifold $\mathbf{C}$ should be close to the RMn of the PSD matrices on the manifold. Iteration and re-iteration of the lifting and projection will result in an estimate very close to the true RMn. In fact, for the measure of d_{R_2}, this is a one-step process as stated in the following [Wong et al. 2017]:

Theorem 1. *For PSD matrices* $\{\mathbf{P}_n, n = 1, \cdots, N\}$*, the RMn according to* d_{R_2} *and the weighted RMn according to* d_{WR_2} *are respectively given by*

$$\mathbf{C}_{\mathrm{R}_2} = \tilde{\mathbf{C}}_{\mathrm{R}_2} \cdot \tilde{\mathbf{C}}_{\mathrm{R}_2}^{H} \tag{3.20a}$$

$$\mathbf{C}_{\mathrm{WR}_2} = \tilde{\mathbf{C}}_{\mathrm{WR}_2} \tilde{\mathbf{C}}_{\mathrm{WR}_2}^{H} \tag{3.20b}$$

where $\tilde{\mathbf{C}}_{\mathrm{R}_2} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{P}_n^{1/2}$,

and $\tilde{\mathbf{C}}_{\mathrm{WR}_2} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{\Omega}^H \mathbf{P}_n \mathbf{\Omega})^{1/2}$ □

Proof of Theorem 1 can be shown readily by substituting d_{R_2} and d_{WR_2} correspondingly into Eq. (3.18) and equating the differentials of the objective functions w.r.t. $\tilde{\mathbf{C}}$ to zero.

Both the EMn and the RMn of PSD matrices have important applications in signal processing. We will employ both in the study of EEG signal classification in the ensuing chapters of the thesis.

Chapter 4

Classification of EEG Signal Features

After features have been selected and extracted from EEG signals, classification begins. Classification algorithms can be divided into different categories based on different perspectives such as linear classifiers, nonlinear classifiers, and combinations of classifiers [Lotte et al. 2007].

The popular linear classifiers used in EEG signal classifications are linear discriminate analysis (LDA) and linear support vector machines (SVM). LDA [Duda et al. n.d.] assumes that the data in each class has normal (Gaussian) distribution all having the same covariance matrix. The separating hyperplane is constructed by seeking the projection that maximizes the distance between the means of two classes and minimizes the variance of interclass. LDA classifier has a very low computational requirement which makes it suitable in online applications [Garrett et al. 2003]. The main drawback of LDA is that it gives poor performance on complex nonlinear EEG data [Garcia et al. 2003]. Linear SVM [Duda et al. n.d.] aims to find a hyperlane that maximizes the margins, i.e., the distance from the nearest training points. Linear SVM has been successfully applied to synchronous brain computer interface (BCI) problems [Garrett et al. 2003]. By using "kernel trick" the linearity restriction can be relaxed so that nonlinear decision boundaries can be created, with only a low increase of the classifier's complexity. The

radial basis function (RBF) SVM also have successful applications in EEG signal classification [Garrett et al. 2003]. SVM has good generalization properties due to the margin maximization and the regularization. It is insensitive to overtraining. It overcomes the problem of "curse-of-dimensionality". The drawback is the low speed of execution.

The nonlinear classifiers mostly used in EEG signal classification are the Nonlinear Bayesian classifiers [Keirn and Aunon 1990]. Another choice is the Hidden Markov model (HMM) classifiers because it is not necessary to extract feature vectors from EEG signals for the classification. HMM has been used successfully in BCI [Obermaier et al. 2001a, Obermaier et al. 2001b] and sleep staging [Doroshenkov et al. 2007].

A neural network can be viewed as universal approximator of continuous functions. Thus, it can produce nonlinear decision boundaries when used in classification [Bishop, Bishop, et al. 1995]. However, the universality makes the classifiers sensitive to overtraining, especially with noisy and non-stationary data. Therefore, one must be careful to select the architecture and regularization [Duin and Tax 2005]. Multilayer perceptron (MLP), together with linear classifiers, are the neural networks mostly used in EEG signal classifications [Hiraiwa et al. 1990, Wang et al. 2004, Balakrishnan and Puthusserypady 2005]. Other neural network architectures have also been applied to EEG signal classifications [Millan et al. 2000].

Our problem of EEG signal classification is a binary hypothesis decision problem: either H_0 or H_1. Thus, we are able to use the simplest binary classifiers. In the following, describe these two methods for our purpose of classification of stroke signals.

4.1 Binary Decision Classifier

4.1.1 Binary hypothesis testing

Binary hypothesis testing involves the decision of the occurrence of one or the other hypothesized events (either H_0 or H_1) from an observable z. The observable z is generated from a source through some probability laws. (Thus, in our project, the source is the patient's brain which generates EEG signals yielding the observable feature $\mathbf{P}_n$.) The observable z falls in to the *observation space* $\mathcal{Z}$ which is demarcated into two non-ovelapping regions $\mathcal{Z}_0$ and $\mathcal{Z}_1$. If z falls into $\mathcal{Z}_0$, we decide that H_0 is true. On the other hand, if z falls into $\mathcal{Z}_1$, we decide that H_1 is true. The binary hypothesis testing rules show us how to optimally demarcate the decision regions under different circumstances. Now, binary hypothesis testing generally may have four possible outcomes:

(i) H_0 true — decide D_0 correct decision

(ii) H_1 true — decide D_1 correct decision

(iii) H_0 true — decide D_1 error (false alarm)

(iv) H_1 true — decide D_0 error (miss)

We assign a cost to each of the possibilities: c_{00}, c_{11}, c_{10} and c_{01} and,

we assume: $c_{10} > c_{00},\ c_{01} > c_{11}$. The most complete decision criterion is the Bayes' decision criterion which minimizes the average cost of the decision making [Wong 2004]. This can be expressed as:

$$\Lambda(z) \triangleq \frac{p(z|H_1)}{p(z|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)(c_{10}-c_{00})}{P(H_1)(c_{01}-c_{11})} \triangleq \eta \tag{4.1}$$

On the left side of Eq. (4.1), $\Lambda(z)$ is called the likelihood ratio which is defined as the ratio of two conditional probability density functions, under the hypotheses of H_1 and H_0. On the right side, we have the known quantities $P(H_0)$ and $P(H_1)$ which are the *a priori* probabilities of the two hypotheses, and the given values of the costs of the decision outcomes. If these values are all known, then the right side is a constant called the *threshold* denoted by η.

In particular:

(i) if $c_{10} - c_{00} = c_{01} - c_{11}$, the Bayes' criterion reduces to the *maximum a posterior* (MAP) criterion.

(ii) if $c_{00} = c_{11} = 0$ and $c_{10} = c_{01} = 1$, then we have

$$\Lambda(z) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)} \qquad \text{(min. error} \equiv \text{MAP)}$$

for which the minimization of the average cost becomes the minimization of the average error (minimum error criterion) which is also MAP.

To apply the Bayes' binary decision criterion to our probelm of EEG signal feature classification, we have to formulate the following likelihood ratio test:

$$\frac{p(\mathbf{P}|H_1)}{p(\mathbf{P}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.2}$$

On the left side, $p(\mathbf{P}|H_1)$ and $p(\mathbf{P}|H_0)$ are respectively the probability density distributions of the PSD matrices of stroked patients and of the healthy participants respectively. On the right side is the threshold which is chosen to minimize the average cost if all the conditions are known.

4.1.2 Distance-from-mean (DFM) binary decision

Unfortunately, we do not have any knowledge of either of the distributions on the left side of Eq. (4.2), therefore, we substitute the probability density distributions with the measures of dissimilarity to the two (healthy and unhealthy) groups of PSD matrices. Distance, as we mentioned in Chapter 3 is a masure of dissimilarity and to measure the distance from a point to a group of PSD matrices, we must choose a group representative. In this case, a suitable representative is the *mean* of the group. On the right side of Eq. (4.2), we do not have any knowledge of the costs of decision outcomes, neither do we have any knowledge of the a priori probabilities of being healthy and having stroke. We then assume that $c_{00} = c_{11} = 0$ and $c_{10} = c_{01} = 1$ and that $P(H_0) = P(H_1)$, which

yields us the decision rule:

$$\frac{d(\mathbf{P},\mathbf{C}_0)}{d(\mathbf{P},\mathbf{C}_1)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{4.3}$$

where d is a distance measure (either Euclidean or Riemannian), and $\mathbf{C}_0$ and $\mathbf{C}_1$ are respectively the means of the PSD Matrices of the healthy and unhealthy participants. Again, we can use either the Euclidean mean or the Riemannian mean for the purpose. We call the testing rule of Eq. (4.3) the Distance-from-mean (DFM) binary decision rule. Eq. (4.3) uses $\mathbf{C}_0$ and $\mathbf{C}_1$ as references for deciding if a testing PSD matrix $\mathbf{P}_{\mathrm{T}}$ is from a healthy person or a stroked patient. On the other hand, we may also translate our decision reference to the origin. In that case, *on average*, we expect the PSD matrix $\mathbf{P}_{\mathrm{T}}$ to have the form,

$$\mathbf{P}_{\mathrm{T}} = \begin{cases} \mathbf{C}_0 & \text{healthy person} \\ \mathbf{C}_1 & \text{stroked patient} \end{cases}$$

Hence, our hypothesis testing rule in this case is:

$$\frac{d((\mathbf{P}-\mathbf{C}_0),\mathbf{0})}{d((\mathbf{P}-\mathbf{C}_1),\mathbf{0})} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{4.4}$$

Eq. (4.4) translates the reference from the means to $\mathbf{0}$. In general, we can use either Eq. (4.3) or Eq. (4.4) for the binary decision test. If Euclidean distance is used, the two equations will yield identical results since there is no distortion in distance if we move the reference. However, if Riemannian distance is used, the distances may have been distorted, and resulting in greater or less accuracies, depending on the distortion.

By choosing different distance measures of d together with the corresponding mean points $\mathbf{C}_0$ and $\mathbf{C}_1$ in the decision rule of Eqs. (4.3) and (4.4), we have different EEG stroke classifiers. Varying the value of the threshold on the right side, the *receiver operation characteristic* (ROC) [Van Trees 2004] of the different classifiers can be obtained and a comparison of the performance can be made.

Clearly, without the knowledge of the density distributions $p(\mathbf{P}|H_1)$ and $p(\mathbf{P}|H_0)$, as well, without the knowledge of $P(H_0)$ and $P(H_1)$, the decision using Eq. (4.3) can only be sub-optimum. However, we can use the *a priori* knowledge of the library of the collected PSD matrices to improve ourresults. This is the method of weighting.

4.1.3 Optimum weighting for binary decision

We have seen in Chapter 3 how the different distance measures can be weighted. The purpose of weighting a distance is to use prior information to highlight certain parts, and deemphasize others, of the feature matrices so as to increase the efficiency of signal processing. For binary signal classification, which is to distinguish one kind of signal feature from another, the optimum weighting matrix should enhance their dissimilarity. We may define similarity between two feature PSD matrices $\mathbf{P}_m$ and $\mathbf{P}_n$ as the amount of correlation such that $\sigma(\mathbf{P}_m, \mathbf{P}_n) = \text{tr}\ (\mathbf{P}_m^H \mathbf{P}_n)$. Suppose for our prior knowledge, we divide our library of collected PSD matrices into two classes: $\mathcal{S}_0$ and $\mathcal{S}_1$, indicating respectively, the group of healthy people and the group of stroked patients, and each having respectively N_0 and N_1 number of epochs. Then, for the purpose of classification, we seek for a weighting matrix which maximizes the correlation between matrices of similar classes and minimizes the correlation between dissimilar classes. In particular, if $\mathbf{C}_{0\text{W}}$ and $\mathbf{C}_{1\text{W}}$ are respectively the weighted Riemannian means of the healthy and stroked PSD matrices, we seek for an optimum $M \times M$ weighting matrix $\mathbf{W} = \mathbf{\Omega}\mathbf{\Omega}^H$ minimizing the following objective function

$$F_{\text{o}}(\mathbf{\Omega}) = \text{tr}\left[\mathbf{C}_{1\text{W}}^{-1}\mathbf{C}_{0\text{W}}\right] \tag{4.5}$$

Since $\text{tr}[\mathbf{A}^{-1}\mathbf{B}] \geq (\text{tr}\mathbf{B})/(\text{tr}\mathbf{A})$ [Wong et al. 2017], we can interpret from the identity $F_{\text{o}}(\mathbf{\Omega}) = \text{tr}[(\mathbf{C}_{1\text{W}}\mathbf{C}_{1\text{W}})^{-1}\mathbf{C}_{1\text{W}}\mathbf{C}_{0\text{W}}]$ that the objective function of Eq.(4.5) is an upper bound of $[\text{tr}(\mathbf{C}_{1\text{W}}\mathbf{C}_{0\text{W}})]/[\text{tr}(\mathbf{C}_{1\text{W}}\mathbf{C}_{1\text{W}})]$, so that minimizing $F_{\text{o}}(\mathbf{\Omega})$ is indeed minimizing the upper bound of the ratio of the correlation between the two means of dissimilar classes to that of similar class.

Since the means for both weighted RD d_{WR_1} and d_{WR_2} are determined by methods which assume the weighting to be fixed, to find the optimum weighting in terms of these means as in Eq. (4.5) necessitates an iterative procedure. In any case, we need the solution of the optimum weighting for fixed means. For that we quote the following theorem [Wong et al. 2017]

Theorem 2. *Suppose we have the following objective function*

$$F_{\mathrm{o}}(\mathbf{\Omega}) \quad = \quad \mathrm{tr}\left[(\mathbf{\Omega}^H \mathbf{\Pi}_1 \mathbf{\Omega})^{-1} (\mathbf{\Omega}^H \mathbf{\Pi}_0 \mathbf{\Omega})\right] \tag{4.6}$$

where $\mathbf{\Pi}_1$ *and* $\mathbf{\Pi}_0$ *are respectively some chosen means among the stroke patients and healthy people PSD matrices. Let* $\{\lambda_1 \geq \cdots \geq \lambda_M\}$ *and* $\{\mathbf{u}_1, \cdots, \mathbf{u}_M\}$ *be respectively the eigenvalues and eigenvectors of* $\mathbf{\Pi}_1^{-1}\mathbf{\Pi}_0$*, then the maximum and minimum values of* F_{o} *are attained when* $\mathbf{\Omega}_{\mathrm{op}}$ *is comprised respectively of the first* K *and the last* K *eigenvectors of* $\{\mathbf{u}_1, \cdots, \mathbf{u}_M\}$. □

The proof of Theorem 2 is shown in [Wong et al. 2017]. ■

Remarks on Theorem 2

1. If we choose $K = M$, then, the resulting weighting matrix $\mathbf{W} = \mathbf{I}$, which does not have any effect. To have the largest possible effect of weighting, we choose $K = M - 1$.

2. For $K = M - 1$, the optimum weighting matrix can yield a minimum $\sum_{i=2}^{M} \lambda_i$ or a maximum $\sum_{i=1}^{M-1} \lambda_i$ for F_{o}. For our purpose, we want to minimize F_{o}. Thus, we choose the eigenvectors corresponding to the K smallest eigenvalues to construct $\mathbf{\Omega}_{\mathrm{op}}$.

3. Theorem 2 provides us with the way to find the optimum weighting matrix for given healthy and stroked means. However, we have to combine this theorem with the algorithms to locate the different means according to weighted d_{WR_1} and d_{WR_2}. This is illustrated in the following.

General algorithm to find optimum weighting matrix for RD:

1. Set accuracy indicators ϵ_0 and ϵ_1.

2. For $i = 0$, set

$$\mathbf{\Pi}_1^{(i)} = \frac{1}{N_1} \sum_{\mathbf{P}_m \in \mathcal{S}_1} \mathbf{P}_m; \quad \mathbf{\Pi}_0^{(i)} = \frac{1}{N_0} \sum_{\mathbf{P}_n \in \mathcal{S}_0} \mathbf{P}_n$$

3. For $K = M$, use Theorem 2 together with $\mathbf{\Pi}_1^{(i)}$ and $\mathbf{\Pi}_0^{(i)}$ to obtain $\mathbf{\Omega}^{(i)}$.

4. Form the weighted healthy PSD group and the weighted stroked PSD group such that

$$\begin{aligned} \mathcal{S}_{0\mathrm{W}}^{(i)} &= \left\{ (\mathbf{\Omega}^{(i)})^H \mathbf{P}_m \mathbf{\Omega}^{(i)} \right\}_{\mathbf{P}_m \in \mathcal{S}_0} \\ \mathcal{S}_{1\mathrm{W}}^{(i)} &= \left\{ (\mathbf{\Omega}^{(i)})^H \mathbf{P}_m \mathbf{\Omega}^{(i)} \right\}_{\mathbf{P}_m \in \mathcal{S}_1} \end{aligned}$$

5. Use the mapping $\mathbf{P} = \tilde{\mathbf{P}}^2$ to lift these weighted PSD matrices to the corresponding isometric Euclidean subspace and, use Theorem 1 to locate the means ($\tilde{\mathbf{C}}_{1\mathrm{W}}^{(i)}$ and $\tilde{\mathbf{C}}_{0\mathrm{W}}^{(i)}$). Project back to the manifold and obtain the new central points, $\mathbf{C}_{1\mathrm{W}}^{(i)}$ and $\mathbf{C}_{0\mathrm{W}}^{(i)}$.

6. Calculate the unweighted healthy and stroked means

$$\begin{aligned} \mathbf{C}_{0\mathrm{R_w}}^{(i)} &= (\mathbf{\Omega}^{(i)})^{-H} \mathbf{C}_{0\mathrm{W}}^{(i)} (\mathbf{\Omega}^{(i)})^{-1} \\ \mathbf{C}_{1\mathrm{R_w}}^{(i)} &= (\mathbf{\Omega}^{(i)})^{-H} \mathbf{C}_{1\mathrm{W}}^{(i)} (\mathbf{\Omega}^{(i)})^{-1} \end{aligned}$$

7. If $d_{\mathrm{R}_2}(\mathbf{C}_{0\mathrm{R_w}}^{(i)}, \mathbf{C}_{0\mathrm{R_w}}^{(i-1)}) < \epsilon_0$ and $d_{\mathrm{R}_2}(\mathbf{C}_{1\mathrm{R_w}}^{(i)}, \mathbf{C}_{1\mathrm{R_w}}^{(i-1)}) < \epsilon_1$, then let $\mathbf{\Omega}_0 = \mathbf{\Omega}^{(i)}$. Since we want to minimize F_{o}, we form $\mathbf{\Omega}_{\mathrm{op}}$ using the last, $M - 1$ columns of $\mathbf{\Omega}_0$ in the way as decribed in Theorem 2. Then exit.

Otherwise, let $\mathbf{\Pi}_0^{(i+1)} = \mathbf{C}_{0\mathrm{R_w}}^{(i)}$ and $\mathbf{\Pi}_1^{(i+1)} = \mathbf{C}_{s\mathrm{R_w}}^{(i)}$. Let $i \to (i+1)$ and go to Step 3. ■

4.2 k-Nearest Neighbor (k-NN) Classification

k-nearest neighbor classifiers are the simplest and among the most effective nonlinear classifiers. The idea is to assign a feature vector to a class according to its nearest neighbors. The neighbors can be feature vectors from the training set if a distance measure is defined between feature vectors [Borisoff et al. 2004], or class prototypes if Mahalanobis distance is used [Cincotti et al. 2003]. The performance of a k-nearest neighbor classifier can be equal to that of a neural network classifier in the automatic scoring of human sleep recordings [Becq et al. 2005].

In this chapter, we introduce classification by k-NN and classification by binary hypothesis testing.

4.2.1 Nearest neighbor classification methods

For our case of EEG signal classification, we take the feature PSD matrix of a test signal epoch not being part of the library, and compare the distance (both Euclidean and Riemannian) of this test feature *Power Spectral Density* (*PSD*) matrix to its k nearest neighbors. Then we assign it to a class according to majority decision among these k neighbor matrices. Fig. 4.1 shows an example of 3-NN and an example of 5-NN in a two-class case. (In our case, the neighbors are the feature *PSD* matrices from the library signal sets.)

We can see that the assignment of the object x may vary with the choice of different values of k, regardless of whether the distributions of the objects are similar or different. However, there is no general rule to choose the best value of k in the k-nearest neighbor algorithm. If the sample size is infinite, the larger k the better is the performance of the k-nearest neighbor classifier. In fact, for infinitely large sample-size, the performance

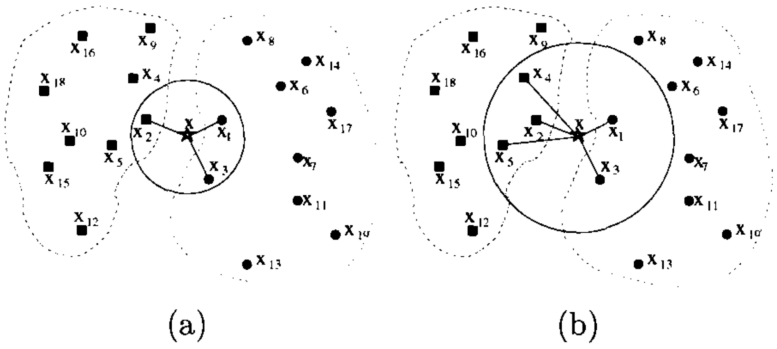

FIGURE 4.1: *k-Nearest Neighbor Decision (a). $k = 3$ (b). $k = 5$*

of the k-nearest neighbor algorithm has been shown to approach the optimum Bayesian classifier with $k \to \infty$ and $k/N \to 0$ (N being the sample size) [Devroye et al. 2013].

In our tests, we first set up a library of epochs of *Electroencephalography* (*EEG*) signals and categorize them into $L = 2$ classes, representing the healthy and the stroked groups. Each epoch of EEG signals has been taken from participants examined by clinical experts and classification agreements have been obtained. Using the procedure described in Section 2.5, the *PSD* matrices of these signal epochs in each of the categories are evaluated at each frequency point within the range $\omega \in [0Hz, 13Hz]$ forming frequency curves (sequences of points), $\{\mathbf{P}_n(\omega), n = 1, \ldots, N\}$, of the N epochs. These are the *PSD* matrix curves to which we apply the k-nearest neighbor algorithm incorporating the Euclidean or Riemannian distances for classification of the *EEG* signals. Since our sample size is finite, we found that, by choosing a small value of k, the results are very satisfactory.

4.2.2 Optimum Riemannian distance weighting for k-NN classification

In the previous chapters, we introduced *Riemannian Distance* (*RD*) for the measure of similarity between two *PSD* matrices and *RD* for weighted *PSD* matrices. In this section, we will obtain the optimum weighting matrix $\mathbf{W}$ in order to enhance the application of this distance for *EEG* classification.

Since the distance measure characterizes the property of the class in which similar data are clustered, the mean-square distance between members of the class is a measure of the size of the cluster so formed. The aim of metric learning is therefore to find the optimum weighting matrix which minimizes the size of the cluster and extract the property of the set in which they are most similar while keeping the dissimilar members at a prescribed distance [Sebestyen 1962]. Conceptually, we can apply this idea to obtain an optimum weighting matrix for the *RD* d_{RW} between two weighted *PSD* matrices [Li and Wong 2013] as follows:

Let $\mathbf{P}_i(\omega)$ and $\mathbf{P}_j(\omega)$, $\omega \in [\omega_{min}, \omega_{max}]$, be two separate samples curves of *PSD* matrices as the frequency ω varies. We say that $\mathbf{P}_i(\omega)$ and $\mathbf{P}_j(\omega)$ are similar if they belong to the same class, and are dissimilar if they belong to different classes. Let $\mathbf{P}_{ik} = \mathbf{P}_i(\omega_k)$ and $\mathbf{P}_{jk} = \mathbf{P}_j(\omega_k)$ represent two separate *PSD* matrices from the two sample curves measured at $\omega = \omega_k$. We denote the sets of similar and dissimilar *PSD* matrices by $\mathcal{S}$ and $\mathcal{D}$ respectively such that the set of pairs of similar *PSD* matrices is $\mathcal{S} = \{(\mathbf{P}_{ik}, \mathbf{P}_{jk}); \mathbf{P}_i(\omega), \mathbf{P}_j(\omega) \in C_l\}$, whereas the set of pairs of dissimilar *PSD* matrices is $\mathcal{D} = \{(\mathbf{P}_{ik}, \mathbf{P}_{jk}); \mathbf{P}_i(\omega) \in C_{l_i}, \mathbf{P}_j(\omega) \in C_{l_j}, l_i \neq l_j\}$. The optimum $M \times M$ weighting matrix $\mathbf{W}$ may be found by maximizing the ratio of the sum of squared interclass distances and the sum of squared of intraclass distances, i.e.,

$$\begin{aligned} \max_{\mathbf{W}} \quad & \sum_{(\mathbf{P}_{m_i}, \mathbf{P}_{n_i}) \in \mathcal{A}_d} d^2_{\mathrm{WR}}(\mathbf{P}_{m_i}, \mathbf{P}_{n_i}) \Big/ \sum_{(\mathbf{P}_{m_i}, \mathbf{P}_{n_i}) \in \mathcal{A}_s} d^2_{\mathrm{WR}}(\mathbf{P}_{m_i}, \mathbf{P}_{n_i}) \\ s.t. \quad & \mathbf{W} = \mathbf{W}^H \succ \mathbf{0} \end{aligned} \tag{4.7}$$

For our case, the total number of pairs of similar and dissimilar *PSD* matrices are respectively given by

$$N_s = \binom{N_{l_p}}{2} + \binom{N_{l_h}}{2}, \quad and \quad N_d = N_{l_p} \cdot N_{l_h}. \tag{4.8}$$

Optimizing of the quantity in (4.7) directly on manifold $\mathcal{M}$ is difficult. However, the

optimization could be performed using inner product metric in an Euclidean space [Li and Wong 2013]. Since we only employ d_{R_2} in this thesis, we will focus on finding the optimum weighting for d_{WR_2}:

Optimum weighting for d_{WR_2}

Let $\mathcal{S}_2 = \{\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}; \mathbf{P}_i(\omega), \mathbf{P}_i(\omega) \in C_l\}$ and $\mathcal{D}_2 = \{(\mathbf{P}_{ik}, \mathbf{P}_{jk}); \mathbf{P}_i(\omega) \in C_(l_i), \mathbf{P}_j(\omega) \in C_{l_j}, l_i \neq l_j\}$. Then, writing

$$\tilde{\mathbf{M}}_{S_2} = \sum_{(\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}) \in \mathcal{S}_2} (\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})(\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})^H \tag{4.9}$$

and

$$\tilde{\mathbf{M}}_{D_2} = \sum_{(\mathbf{P}_{ik}^{1/2}, \mathbf{P}_{jk}^{1/2}) \in \mathcal{D}_2} (\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})(\mathbf{P}_{ik}^{1/2} - \mathbf{P}_{jk}^{1/2})^H \tag{4.10}$$

The optimization problem in Eq. (4.7) becomes

$$\max_{\mathbf{\Omega}} \ \mathrm{tr}\left(\mathbf{\Omega}^H \tilde{\mathbf{M}}_{\mathcal{D}1} \mathbf{\Omega}\right) \Big/ \mathrm{tr}\left(\mathbf{\Omega}^H \tilde{\mathbf{M}}_{\mathcal{S}1} \mathbf{\Omega}\right), \ \text{s.t. } \mathbf{\Omega}\mathbf{\Omega}^H \succ 0 \tag{4.11}$$

A globally optimum solution of the above may still be difficult to find due to the non-convexity of the problem. Thus, we seek to solve an approximation to the problem, such that

$$\max_{\mathbf{\Omega}} \ \mathrm{tr}\left[(\mathbf{\Omega}^H \tilde{\mathbf{M}}_{\mathcal{S}1} \mathbf{\Omega})^{-1} \mathbf{\Omega}^H \tilde{\mathbf{M}}_{\mathcal{D}1} \mathbf{\Omega}\right], \ \text{s.t. } \mathbf{\Omega}\mathbf{\Omega}^H \succ \mathbf{0} \tag{4.12}$$

The solution of Eq. (4.12) is presented in [Li and Wong 2013], such that the optimum weight is given by

$$\mathbf{\Omega}_{op2} = [\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k]^T \tag{4.13}$$

where $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k$ are the orthonormal eigenvectors corresponding to the eigenvalues $\tilde{\lambda}_1 \leq \cdots \leq \tilde{\lambda}_K$ of $\tilde{\mathbf{M}}_{S_2}^{-1}\tilde{\mathbf{M}}_{D_2}$. Thus, the optimum weighting matrix $\mathbf{W}_{op2}$ is given by

$$\mathbf{W}_{op2} = \mathbf{\Omega}_{op2}\mathbf{\Omega}_{op2}^{H} \tag{4.14}$$

4.2.3 Summary of k-NN classification procedure

In the following, we summarize the classification of *EEG* signals using the k-nearest neighbor algorithm incorporating the unweighted Riemannian distance d_{R_2} and weighted Riemannian distance d_{WR_2}. (For classification using the k-nearest neighbor with other weighted or unweighted distances, the procedure will be identical):

1. With all the *PSD* matrices of *EEG* signal epochs of the L label in the data set, the optimum weighing matrix $\mathbf{W}$ of the similarity/dissimilarity is evaluated for $d_{R_{2W}}$.

2. For the *PSD* matrix curve $\mathbf{P}_0(\omega)$ of a test *EEG* signal, we calculate the dissimilarity measures $\{d_{ni} = d_{WR_2}(P_0(\omega_i), P_n(\omega_i)), n = 1, \ldots, N\}$ at each frequency ω_i according to (3.13), and then calculate the total distance d_n between the two curves according to (3.17). For a chosen value of k, the k nearest neighbors of the test signal $\mathbf{P}_0(\omega)$ (k *PSD* matrices at same ω having shortest weighted distances from $\mathbf{P}_0(\omega)$) are then identified.

3. $\mathbf{P}_0(\omega)$ is then assigned to class $\mathcal{C}_{l_0}$ if $l_0 = maj\{l_p, l_h\}$ where $\{l_p, l_h\}$ are the class labels of the k-nearest neighbors of $\mathbf{P}_0(\omega)$ among the members of the data set, and $maj(\cdot)$ denotes the majority vote function, i.e., its value is the element which occurs most in $\{l_p, l_h\}$.

4.3 Q-fold cross-validation method

In this section, we will discuss the method for examine the performance of our classification methods. Ideally, the performance accuracy of our *EEG* classification algorithm

should be measured in terms of its probability of error which necessitates the knowledge of the ground truth of if patient has stroke or not. However, since the ground truth of health state of a patient measured from the signal epoch is not really known, we will therefore treat the library of signal epochs classified by clinical experts as the ground truth. From the library of collected signal epochs, we will randomly select some as training signals and some as test signals so that the validation of our classification methods is carried out as follows:

1. For each of the classes C_l, $l = \{0, 1\}$, containing N_l feature *PSD* matrix curves (being functions of ω) of the patients or healthy person, we randomly choose N_{lT} matrix curves as the test set and the rest $(N_l - N_{lT})$ as the training (library) set.

2. As described in the previous sections, for all the two classes of *EEG* signal, the weighting matrix $\mathbf{W}$ is first evaluated using the training sets, each containing $(N_l - N_{lT})$ selected feature matrix curves. The evaluation of the weighting matrix and the weighted mean follows the procedure of Section 4.1.3 if DFM binary decision is used, wheresas if k-NN classification is employed, the weighting matrix is evaluated as described in Section 4.2.2. For each member matrix curve of the test sets, the dissimilarity measures from the library sets are calculated and its classification is carried out according to either the DFM decision rule or the k-NN classification rule.

3. The above steps are repeated Q times (Q-fold cross-validation), each time choosing different sets of training and test feature matrix curves in C_l. The probability of correct classification for each class can then be estimated by $\hat{P}_{cl} = \frac{1}{Q}\sum_{q=1}^{Q}\hat{P}_{clq}$ where $\hat{P}_{clq}$ denotes the estimated probability of correct classification of class l at the qth trial, $q = 1, \ldots, Q$, i.e., $\hat{P}_{clq} = \frac{N_{lc}}{N_{lT}}$ with N_{lc} and N_{lT} being the number of correct classification and total number of members in class C_l at the qth trial.

Chapter 5

Experimental Verifications

To evaluate the performance of the different classification methods using different signal features, we first introduce some terminology borrowed from radar technology.

- *False alarm* – This is an error made by classifying a healthy person as a stroked patient,
- *Miss* – This is an error made by classifying a stroked patient as a healthy person.
- *Detection* – A classification is called detection when a stroked patient is correctly classified.
- Receiver Operating Characteristic (ROC) – This is a curve which shows the probability of false alarm against the probability of detection of the classifier.

In this chapter, we evaluate the performance of different feature extraction methods together with different classifiers using the following values:

1. **Overall accuracy** – This is calculated by the total correct classification number divided by total test number.
2. **Individual class accuracy** – These two indicate the accuracies for healthy person and stroked patient classification.
3. **Confusion matrix** – This is a 2×2 matrix containing the numbers $N_{00}, N_{01}, N_{10}, N_{11}$, where N_{00} denotes the number of correctly classified healthy persons, N_{01} denotes

the number of flase alarms, N_{10} denotes the number of misses, and N_{11} denotes the number of correct detections.

4. **Receiver Operating Charateristic (ROC)** – We plot the ROC curves of the different classification methods in which their probabilities of false alarm and miss are approximated by their experimental rates of false alarm and miss. The *Area Under the Receiver Operating Characteristic Curve* (AUC) is an indiocator of the the godness of performance. Specifically, the closer is AUC to 1, the better performance of a classification method.

We now perform some tests using the collected *Electroencephalography* (EEG) signal to validate our classification algorithm employing the Riemannian distance developed in Chapter 3 and Chapter 4. The test results are based on the data collected from 45 person in which 23 are healthy persons and 22 are stroke patients. For each person, we collect the multichannel signals recorded from channels C_3, C_4, O_1, O_2 illustrated in Fig. 2.1. As described in the previous chapters, for our validation tests, the raw EEG recordings were first pre-processed by removing the DC values (referencing), and the frequency components of the signals were kept to within the range of $0.5 - 13Hz$ by using a bandpass filter. We splitted the recording length to $30s$ epochs. Each epoch was examined and labelled as either patient(p) or healthy(h) according to his/her health record. Thus, we establish a library of two categories of EEG records. The *Power Spectral Density* (PSD) matrices of each epoch were then evaluated by the Nuttall-Strand algorithm [Strand 1977b, Nuttall 1976]. In each trial, we randomly choose 125 PSD matrices from each category as test signals while the remaining $(2494 - 125)$ PSD matrices form the training data set so that the total number of the training signal feature in each trial is 2369.

The following are examples of the tests of the effectiveness of various dissimilarity measures in the classification of EEG signals we carried out under different environments.

5.1 Distance comparison binary hypothesis testing result

Now we examine the performance of the hypothesis testing rule for different distance measures by computer simulations:

- First, from our library of collected *PSD* matrices from patient's and healthy person's *EEG* signal, we calculate an optimum weighting matrix for the *Riemannian Distance* (RD).
- Second, we substitute one by one the various distance measures d_E, d_{R_2} and the weighted version d_{WR_2}, together with the corresponding means into (4.3) and (4.4) respectively for binary decision according to the distance from the mean, and according to the distance from the origin.
- Finally, for a range of thresholds we calculate the false alarm rate, the missing rate, the rate of detection and record the overall accuracy, AUC, accuracy for each class, and the confusion matrices.

TABLE 5.1: Confusion matrix for Euclidean metric

		All-bands		Delta band		Theta band		Alpha band	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1242	126	1238	130	1313	55	223	1145
	P	771	355	767	359	889	237	105	1021

TABLE 5.2: Accuracy of Euclidean distance

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.555	0.64	0.908	0.315
Delta band	0.561	0.64	0.905	0.319
Theta band	0.533	0.621	0.96	0.21
Alpha band	0.612	0.499	0.163	0.907

5.1.1 Example 5.1.1

First, we exam our classification method with *PSD* matrices by Euclidean metric. The confusion matrix of Euclidean metric shown in Table 5.1, and accuracy shown in Table 5.2. The accuracy of all-band test as same as the accuracy of *delta* band test outperform 2% to 14% the accuracy of *theta* and *alpha* band test. According to the confusion matrix, the number of error from missing is triple of the number of error from false alarm. Similar result also shown in the Table 5.2 where accuracy of healthy person class outperforms around 60% accuracy of patient class except *alpha* band test. In the *alpha* band test, the performance has opposite result of other three tests where accuracy of healthy person class only achieved 0.163% accuracy.

TABLE 5.3: Confusion matrix for *RD*

		All-bands		Delta band		Theta band		Alpha band	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1138	230	1141	227	1154	214	1023	345
	P	582	544	572	554	646	480	624	504

TABLE 5.4: Accuracy for *RD*

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.688	0.674	0.832	0.483
Delta band	0.684	0.68	0.834	0.492
Theta band	0.64	0.66	0.844	0.426
Alpha band	0.607	0.611	0.748	0.446

5.1.2 Example 5.1.2

Second, we exam our classification method with *PSD* matrices by *RD*. The confusion matrix of *RD* shown in Table 5.3. The accuracy shown in Table 5.4. Similar to example 5.1.1, the accuracy and *AUC* for all-band test and *delta* test outperform 2% to 8% *theta* test and *alpha* test. Comparing Table 5.4 to Table 5.2, clearly, the *RD* outperforms 3% to 11% Euclidean distance in overall accuracy of classification. The imbalance performance between healthy person class and patient class is still obvious, but the difference between accuracy of healthy person class and patient class reduced from $0.58 - 0.74$ (Table 5.2) to $0.3 - 0.41$ (Table 5.4).

TABLE 5.5: Confusion matrix for *RD* with zero shift

		All-bands		Delta band		Theta band		Alpha band	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1137	231	1130	238	1175	193	995	373
	P	582	544	562	564	657	469	625	501

TABLE 5.6: Accuracy for *RD* with zero shift

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.643	0.674	0.831	0.483
Delta band	0.647	0.679	0.826	0.501
Theta band	0.595	0.659	0.859	0.417
Alpha band	0.543	0.6	0.727	0.445

5.1.3 Example 5.1.3

Third, we exam our classification method with *PSD* matrices by *RD* with zero shift. The confusion matrix of *RD* with zero shift shown in Table 5.5. Accuracy shown in Table 5.6. Comparing Table 5.6 and Table 5.4, the effect of zero shift in accuracy of classification is marginal while reduce *AUC* 4% to 6% compared to example 5.1.2. The effect of zero shift for classifier in *Receiver Operating Characteristic* (*ROC*) curve will be shown in section 5.1.6.

TABLE 5.7: Confusion matrix for weighted *RD*

		All-bands		Delta band		Theta band		Alpha band	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1133	235	1134	234	1155	213	1019	349
	P	576	550	564	562	645	481	627	499

TABLE 5.8: Accuracy for weighted *RD*

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.674	0.675	0.828	0.488
Delta band	0.669	0.68	0.829	0.499
Theta band	0.628	0.656	0.844	0.427
Alpha band	0.582	0.609	0.745	0.443

5.1.4 Example 5.1.4

Then, we exam our classification method with *PSD* matrices by weighted *RD*. The confusion matrix of weighted *RD* shown in Table 5.7. Accuracy shown in Table 5.8. Comparing Table 5.7 with Table 5.4, the weighting effect in accuracy is marginal. The effect of weighting in *ROC* curve will be shown in section 5.1.6.

TABLE 5.9: Confusion matrix for weighted *RD* with zero shift

		All-bands		Delta band		Theta band		Alpha band	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1144	224	1140	228	1183	185	1004	364
	P	585	541	567	559	661	465	621	505

TABLE 5.10: Accuracy for weighted *RD* with zero shift

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.647	0.676	0.836	0.48
Delta band	0.651	0.681	0.833	0.496
Theta band	0.599	0.661	0.865	0.413
Alpha band	0.54	0.605	0.734	0.448

5.1.5 Example 5.1.5

Then, we exam our classification method with *PSD* matrices by weighted *RD* with zero shift. The confusion matrix of weighted *RD* with zero shift shown in Table 5.9. Accuracy shown in Table 5.10. Similar to unweighted *RD* shown in example 5.1.4, the effect of zero shift in accuracy of classification for weighted *RD* is marginal while reduce *AUC* $1\% - 4\%$ from weighted *RD* without zero shift. Comparing example 5.1.5 with example 5.1.4, the weighting effect is marginal in accuracy and *AUC*.

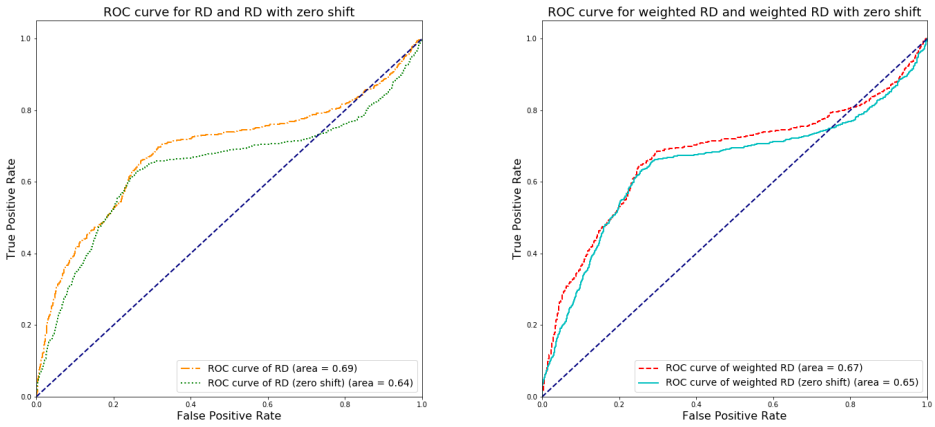

FIGURE 5.1: The effect of zero shift on *ROC* curve

5.1.6 *ROC* comparison for binary hypothesis analysis

In order to compare the performance of different distance metrics for binary hypothesis testing, the *ROC* curves for different pairs of distance metrics are plotted in this section.

The effect of zero shift

The effect of zero shift on *ROC* curves shown in Fig.5.1. The left figure shown the zero shift effect on unweighted *RD*. Clearly, the *ROC* of *RD* is above the *ROC* of *RD* with zero shift. The right figure shown the zero shift effect on weighted *RD*. The *ROC* of weighted *RD* is above the *ROC* of unweighted *RD* but with less difference compared to unweighted *RD*.

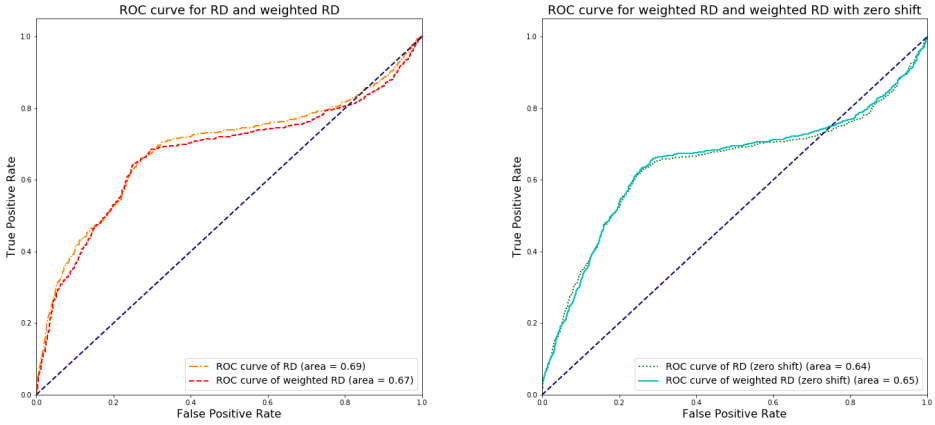

FIGURE 5.2: The effect of optimum weighting on *ROC* curve

The effect of optimum weighting

The effect of optimum weighting on *ROC* curve shown in Fig.5.2. Similar to the conclusion of section 5.1.4 and 5.1.5, the effect of weighting is marginal for both *RD* and zero shift *RD*.

The comparison of different metrics for *ROC* curve

The comparison of different metrics for ROC curve shown in Fig.5.3. Clearly, the ROC curves of RD and weighted RD are above the ROC curve of Euclidean distance. There is cross point at $P_{fp} = 0.1$ where P_{fp} is the false positive rate for the ROC of RD and weighted RD with zero shift. Before the cross point, the ROC of Euclidean distance is above the other two while lower after the cross point.

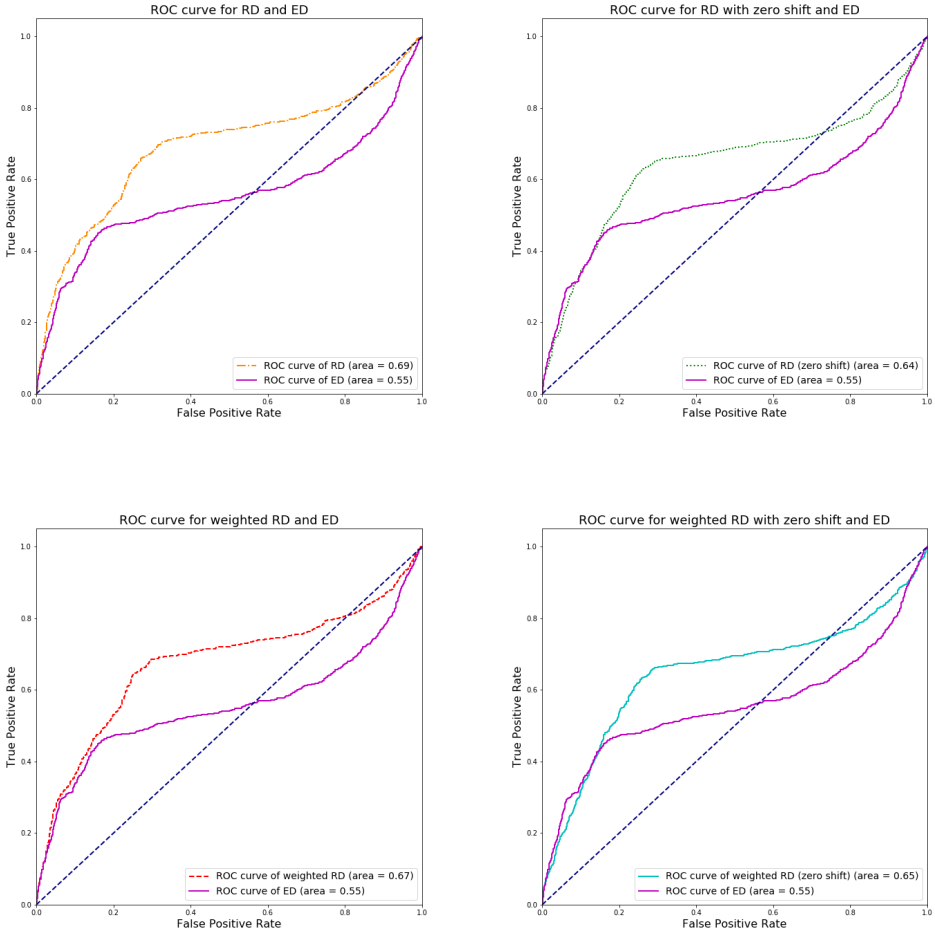

FIGURE 5.3: The comparison of different metrics for *ROC* curve

5.2 *K-Nearest Neighbors* (*KNN*) validation test result

Similar to the section 5.1, we examine the performance of our classification algorithm using the Euclidean distance, the Riemannian distance, and the weighted Riemannian distance and applied on all $delta, theta, alpha$ bands and each sub-band separately by *KNN* classifier. Our experiments are carried out in the same way as in Examples 5.2, that $N_{lT} = 125$ and we employ the parameter $k = 5$ for the nearest neighbor tests. Each test is repeated $Q = 20$ times. Riemannian distances, weighted Riemannian distance and and applied on all $delta, theta, alpha$ bands and each sub-band separately by *KNN* classifier. Our experiments are carried out in the same way that $N_{lT} = 125$ and we employ the parameter $k = 5$ for the nearest neighbor tests. Each test is repeated $Q = 20$ times.

TABLE 5.11: Accuracy of *KNN* for Euclidean Distance

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.9647	0.9102	0.9181	0.9005
Delta band	0.9282	0.8597	0.9028	0.8073
Theta band	0.9433	0.8797	0.9152	0.8366
Alpha band	0.9193	0.8549	0.8735	0.8321

TABLE 5.12: Confusion Matrix of *KNN* for Euclidean Distance

		All-band		Delta		Theta		Alpha	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1256	112	1235	133	1252	116	1195	173
	P	110	1016	217	909	184	942	189	937

5.2.1 Example 5.2.1

In this example, we examine the performance of *KNN* classifier with Euclidean metric. Table 5.11 shows the accuracy of *KNN* with Euclidean metric. We can observe that all-band test obtained 4% to 6% higher accuracy than other sub-band. Similar to binary hypothesis testing, accuracy of healthy person is higher than the accuracy of patient. Confusion matrix in Table 5.12 shows that the number of error from all-band test is less than sub-band test. However, compared to binary hypothesis test, k-NN method yields dramatically higher accuracy. We also observe that here, *theta* band test outperforms other two sub-bands.

TABLE 5.13: Accuracy of *KNN* for *RD*

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.9745	0.9254	0.931	0.9192
Delta band	0.9467	0.8865	0.9269	0.8375
Theta band	0.9541	0.8994	0.9364	0.8544
Alpha band	0.9313	0.8677	0.8823	0.8499

TABLE 5.14: Confusion Matrix of *KNN* for *RD*

		All-band		Delta		Theta		Alpha	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1273	95	1268	100	1281	87	1207	161
	P	91	1035	183	943	164	962	169	957

5.2.2 Example 5.2.2

Example 5.2.2 examines the performance of *KNN* classifier with *RD*. Similar to example 5.2.1, all-band test outperforms 3% to 6% sub-band test in accuracy of classification and *theta* band test achieved the best performance among all sub-band test. The imbalance of accuracy and number of error between healthy person and patient is still obvious.

Comparing example 5.2.1 and example 5.2.2, clearly, *RD* outperforms Euclidean distance by 1% to 3% in accuracy of classification for either all-band test or sub-band test. In comparison to binary decision testing, k-NN also offers dramatically improved results.

TABLE 5.15: Accuracy of *KNN* for weighted *RD*

	AUC	Accuracy	Accuracy (H)	Accuracy (P)
All-band	0.9718	0.9226	0.9247	0.92
Delta band	0.9351	0.8693	0.8947	0.8384
Theta band	0.9397	0.8797	0.9086	0.8446
Alpha band	0.9123	0.8444	0.8545	0.8321

TABLE 5.16: Confusion Matrix of *KNN* for weighted *RD*

		All-band		Delta		Theta		Alpha	
		Predict							
		H	P	H	P	H	P	H	P
True	H	1265	103	1224	144	1243	125	1169	199
	P	90	1036	182	944	175	951	189	937

5.2.3 Example 5.2.3

In this example, we applied the optimum weighting to *RD*. We have simiar observations for all-band and sub-band tests to those in Example 5.2.1 and Example 5.2.2. While the k-NN method with weighted RD improves dramatically in performance from the corresponding test of binary decision, it does not improve on the k-NN method with unweighted RD. This is a little surprising, and we have not been able to find out the reason yet. However, the imbalance performance between two classes in accuracy and number of error has been reduced.

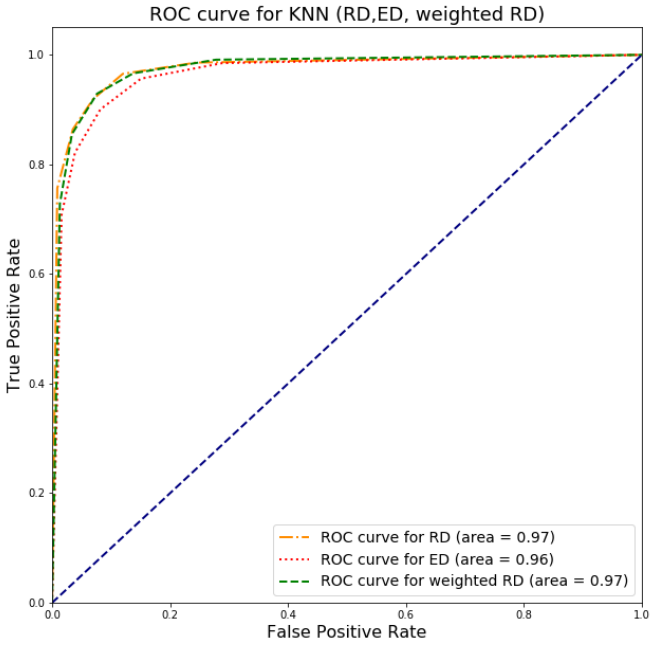

FIGURE 5.4: *ROC* curves for *KNN* classifier

5.2.4 *ROC* result for *KNN*

In order to compare the performance of three distance metric, the *ROC* curves of Euclidean distance, *RD* and weighted *RD* shown in figure 5.4. Clearly, weighted *RD* and *RD* outperform Euclidean distance. The ROC curves for RD and weighted RD almost completely overlap, showing very similar performance.

5.3 Discussion on the Performance of the Different Classification Methods

From our simulations of detection performance, we can observe that different classification methods with different processing result in different performance in the classification of EEG signals. In this section, we try to look into the effectiveness of these different methods and perhaps give some reasons for such results.

We start by reviewing the optimum binary decision method given by Eq. (4.2) in Chapter 4, and is re-written below for reference here.

$$\frac{p(\mathbf{P}|H_1)}{p(\mathbf{P}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5.1}$$

On the left side of this equation of optimum binary hypothesis testing, we have the likelihood ratio which is the ratio of the probability density fucntions of the *obvservable* $\mathbf{P}$ for H_1 and H_0. This is the ratio that we should compare to the threshold to achieve optimum bianry decision. The threshold η on the right is a constant chosen to be best suited to the background knowledge (such as cost of individual decisions, a priori probabilities events) of the environment. If we vary the value of the threshold, and evaluate the probabilities of false alarm and detection, we can examine the performance of this optimum binary decion rule in the form of the ROC. To carry out this optimum binary decision rule, we must therefore have the knowledge of the probability density functions of the observable under both H_1 and H_0.

Now, in our EEG classification experiments, we do not have the knowledge of the probability density functions. Let us see how we try to approach the problem in different ways:

1. Distance Comparison Binary Decision Methods: All the distance comparison methods formulate the decision rule using Eq. (4.3), which is re-written below for

convenience of reference

$$\frac{d(\mathbf{P},\mathbf{C}_0)}{d(\mathbf{P},\mathbf{C}_1)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 \tag{5.2}$$

On the left side of Eq. (5.2), we have a ratio of two distances of the observable measured from the two reference points. The distances can be measured in different metrics, ED, RD, or weighted RD. However, such a distance ratio does not take into account of the distribution of the features, or rather, such a ratio assumes the distribution of both the healthy features and the stroked features are uniformly distributed over the region. On the right of Eq. (5.2) is the threshold, which is set to unity for equal weight, but can be varied to observe the performance of the paricular classifier. Since the distance comparison rule does not account for the local likelihood of the observable being tested, hence all of these methods do not give satisfactory results (see ROC curves Figs. 5.1, 5.2, and 5.3) in the EEG classification experiments. For such distance comparison tests, we also note the following:

- Theoretically, all continuous Likelihood Ratio Test have *ROC* that are concave downwards and are above the line $P_D = P_F$ where P_D is the probability of detection (True Positive Rate) and P_F is the probability of False Alarm (False Positive Rate). However, the ROC curves we obtained in distance comparison tests are not strictly concave (see Figs. 5.1 to 5.3). This may be due to the fact that we used distance ratio $\frac{d_H}{d_H+d_P}$ to represent the probability of stoke detection (P_p) which may not represent accurately the true distributions. Also, if the distributions of the feature matrices are multi-modal or discontinuous, the *ROC* curves for binary hypothesis are not always concave.

- The RD is a more accurate measure of the distance on the PSD manifold than the ED, therefore, for the distance comparison classification, better results can be expected (see Figs. 5.3).

- In an attempt to yield better results from the use of RD, we tried to weight the

feature PSD matrices by contracting the distance betweem similar features and expand the distance between dissimilar features. This, in many cases would help to yield improved classifications [Li et al. 2012, Wong et al. 2017]. In our case the improvement is insignificant (see Figs. 5.2). We suspect that the two distributions of healthy features and stroked features have very significant overlaps and thus contracting distance of like groups and expanding distance of unlike groups cannot be effectively carried out.

- We also attempt to shift the distance to the origin reference. For ED this generally have no effect because the measure of distance is everywhere the same. However, RD have different values measure at different locations. In some cases such as in sonar and MIMO communications [Wong et al. 2017, Han et al. 2017], such a shift of reference warps the distance measure for RD and may yield better results. Unfortunately, in our experiments, we find that the shifting of reference yields lower accuracies (see Figs.5.1). This, we attribute to the nature of the warping of the distance. Depending on the distributions of the feature PSD matrices, the warping of the distances may also turn out to be unfavourable to the classification.

2. k-NN Classification Method: The k-NN method measures the distances between the feature $\mathbf{P}_0$ under test to all the features in the library set and selects the k nearest neighbours of $\mathbf{P}_0$ for examination. If the majority of the k neighbours are healthy features, then $\mathbf{P}_0$ is decided to be healthy. Otherwise, $\mathbf{P}_0$ is classified as stroked. The decision according to the majority of the neighbours is in fact an estimation of the *probabilty distributions* of the two groups in the *neighbourhood* of $\mathbf{P}_0$. This majority decision rule can be interpreted as:

$$P(\mathbf{P}_0 \in \mathcal{P}) \underset{H_0}{\overset{H_1}{\gtrless}} P(\mathbf{P}_0 \in \mathcal{H}) \tag{5.3}$$

where $\mathcal{P}$ and $\mathcal{H}$ respectively denote the stroke patient and the healthy persons groups. Thus, the k-NN classification is an estimation of the likelihood ratio local

to $\mathbf{P}_0$ and compare it to the set threshold. This is certainly a vast improvement from the Distance Comparison methods which assume the distributions of the two groups to be uniform. Hence the classification results of the k-NN methods are all far superior to the Distance Comparison methods (see Table 5.11 to 5.16). In applying the k-NN classification methods, we observe the following:

- All the ROC curves are concave (see Fig. 5.4). This shows that the statement above that k-NN is estimating the local likelihood ratio is correct.
- The use of RD and weighted RD improves on the performance of classification than the use of ED (see Fig. 5.4).
- In all our trials, we use $k = 5$ since this is a convenient and small number to use. The k-NN algorithm have been shown to approach the optimum Bayesian classifier with $k \to \infty$ and $k/N \to 0$ where N is the sample size [Sebestyen 1962]. We have not attempted to find an optimum k since k_{op} is not well defined.

3. In all tests, we observe that the features generated from the data in the all-band test yield the highest accuracies and theta-band yield the highest accuracy among sub-band test.

4. In this thesis, the weighting effect didn't improve the performance in accuracy of classification or number of error, which is different to the theoretical proof. However, based on the observation (see Table 5.13 to 5.16), the weighting effect reduced the imbalance performance in accuracy of classification and error of number.

Chapter 6

Conclusion

6.1 Summary of the thesis

In this thesis, we examined the use of the *Power Spectral Density* (*PSD*) matrix as the feature of *Electroencephalography* (*EEG*) signal classification and demonstrated the advantage of applying *PSD* matrix in the ischemic brain stroke detection by *EEG* signals. We began by introducing *EEG* signal, background knowledge of ischemic brain stroke and commonly used stroke detection methods. Reasoning that the wPSD marices are structurally constrined forming a manifold in the signal space, we suggested the use of RD for the measurement of similarity/dissimilarity between different class of *EEG* signals. We employed the closed-form expressions of the RD and weighted RD developed by Li and Wong. After feature extraction, two classification methods, binary hypothesis testing and *K-Nearest Neighbors* (*KNN*), were applied to detect ischemic brain stroke by extracted *PSD* matrices. In binary hypothesis testing, an iterative algorithm developed by [Wong et al. 2016] was used to locate Riemannian Mean. In order to maximize the correlation between stroke patients' feature and minimize the correlation between stroke patients' and healthy person's feature, the concept of the optimum weighting matrix for *Riemannian Distance* (*RD*) was applied. However, since we use distance instead of distribution of *EEG* signals to represent likelihood ratio, the performance of binary hypothesis is not satisfatory. In the *KNN* test, we evaluate similarity/dissimilarity by

different distance metrics including the commonly used Euclidean metric and Riemannian distance. Furthermore, a pair-wise optimum weighting algorithm is applied to minimize the similarity among the intraclass signals and maximize the similarity among the interclass signals. The simulation results show that using the *PSD* matrices as feature combined with *KNN* classifier for stroke detection provides superior performance compared to using classical power spectrum. Furthermore, the use Reimannian metric as a similarity measure yields hugher accuracies than using Euclidean metric. The use of optimum weighting reduced the difference of the accuracy for each class due to imbalance dataset. In summary, although the stroke detection on *PSD* manifold requires special considerations in distance measures and various algorithms to facilitate the processing, the performance of detection has been significantly improved.

6.2 Future work

There are issuessrising from the research which are worth pursuing. These may be proposed for future work:

1. Considering the performance of two classifiers, the accuracy of sub-band are different and are consistent for each classifier. Therefore, we can raise the following question: “how to apply weighting of frequency band to enhance the accuracy of the classification?"

2. The aim of this research is to detect ischemic brain stroke by *EEG* signal and prevent late diagnosis. However, we have not distinguished the location of the the stroke. Since the PSD matrix mesures the cross-power of the EEG signal emitted from the two hemispheres of the brain, it seems reasonable to assume that it contains sufficient information for locating on which side of the brain the damage occurs. If our library consists of EEG samples which have this information of stroke location, Therefore, we can attempt to use k-NN methods with RD measures to

sub-classify the stroke features into right and left brain hemisphere damages and see how effective it might be.

Bibliography

Anderson, C. W., Knight, J. N., O'Connor, T., Kirby, M. J., and Sokolov, A. (2006). Geometric subspace methods and time-delay embedding for EEG artifact removal and classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14(2), 142–146.

Arnaudon, M., Barbaresco, F., and Yang, L. (2013). Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing* 7(4).

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 56(2), 411–421.

Balakrishnan, D. and Puthusserypady, S. (2005). Multilayer perceptrons for the classification of brain computer interface data. In: *Bioengineering Conference, 2005. Proceedings of the IEEE 31st Annual Northeast*. IEEE, 118–119.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering* 59(4), 920–928.

Barnett, V. and Lewis, T. (1974). *Outliers in statistical data*. Wiley.

Becq, G., Charbonnier, S., Chapotot, F., Buguet, A., Bourdon, L., and Baconnier, P. (2005). Comparison between five classifiers for automatic scoring of human sleep recordings. In: *Classification and Clustering for Knowledge Discovery*. Springer, 113–127.

Bengtsson, I. and Życzkowski, K. (2017). *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press.

Besse, A. L. (2007). *Einstein manifolds*. Springer Science & Business Media.

Bhatia, R. (2009). *Positive definite matrices*. Vol. 24. Princeton university press.

Bianchi, G. and Sorrentino, R. (2007). *Electronic filter simulation & design*. McGraw-Hill New York.

Bishop, C., Bishop, C. M., et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Borisoff, J. F., Mason, S. G., Bashashati, A., and Birch, G. E. (2004). Brain-computer interface design for asynchronous control applications: improvements to the LF-ASD asynchronous brain switch. *IEEE Transactions on Biomedical Engineering* 51(6), 985–992.

Cincotti, F., Scipione, A., Timperi, A., Mattia, D., Marciani, A., Millan, J., Salinari, S., Bianchi, L., and Bablioni, F. (2003). Comparison of different feature classifiers for brain computer interfaces. In: *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*. IEEE, 645–647.

Ciochina, D., Pesavento, M., and Wong, K. M. (2013). Worst case robust downlink beamforming on the Riemannian manifold. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 3801–3805.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media.

Doroshenkov, L., Konyshev, V., and Selishchev, S. (2007). Classification of human sleep stages based on EEG processing using hidden Markov models. *Biomedical Engineering* 41(1), 25–28.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Recognition. 2001.*

Duin, R. P. and Tax, D. (2005). Statistical pattern recognition. In: *Handbook of Pattern Recognition and Computer Vision*. World Scientific, 3–24.

Finnigan, S., Wong, A., and Read, S. (2016). Defining abnormal slow EEG activity in acute ischaemic stroke: Delta/alpha ratio as an optimal QEEG index. *Clinical Neurophysiology* 127(2), 1452–1459.

Foreman, B. and Claassen, J. (2012). Quantitative EEG for the detection of brain ischemia. *Critical care* 16(2), 216.

Franks, L. E. (1969). Signal theory.

Garcia, G. N., Ebrahimi, T., and Vesin, J.-M. (2003). Support vector EEG classification in the Fourier and time-frequency correlation domains. In: *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*. IEEE, 591–594.

Garrett, D., Peterson, D. A., Anderson, C. W., and Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on neural systems and rehabilitation engineering* 11(2), 141–144.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*. Vol. 3. JHU Press.

Gratton, G., Coles, M. G., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and clinical neurophysiology* 55(4), 468–484.

Graybill, F. A. (1983). Matrices with applications in statistics.

Guerrero-Mosquera, C. and Vazquez, A. N. (2009). Automatic removal of ocular artifacts from EEG data using adaptive filtering and independent component analysis. In: *Signal Processing Conference, 2009 17th European*. IEEE, 2317–2321.

Han, G., Zhang, J.-K., Dong, Z., and Mu, X. (2017). Uniquely factorable space-time modulation for two-user uplink massive MIMO systems. In: *Signal Processing Advances in Wireless Communications (SPAWC), 2017 IEEE 18th International Workshop on*. IEEE, 1–5.

Haykin, S. S. (2008). *Adaptive filter theory*. Pearson Education India.

Hiraiwa, A., Shimohara, K., and Tokunaga, Y. (1990). EEG topography recognition by neural networks. *IEEE Engineering in Medicine and Biology Magazine* 9(3), 39–42.

Horn, R. A., Horn, R. A., and Johnson, C. R. (1990). *Matrix analysis*. Cambridge university press.

Jost, J. and Jost, J. (2008). *Riemannian geometry and geometric analysis*. Vol. 42005. Springer.

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., and Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications* 316(1-4), 87–114.

Keirn, Z. A. and Aunon, J. I. (1990). A new mode of communication between man and his surroundings. *IEEE transactions on biomedical engineering* 37(12), 1209–1214.

Kendall, M. G. et al. (1946). The advanced theory of statistics. *The advanced theory of statistics.* (2nd Ed).

Larsen, D. (2015). *A Book about the Film Monty Python and the Holy Grail: All the References from African Swallows to Zoot*. Rowman & Littlefield.

Li, Y. and Wong, K. M. (2013). Riemannian distances for signal classification by power spectral density. *IEEE Journal of Selected Topics in Signal Processing* 7(4), 655–669.

Li, Y., Wong, K. M., and Bruin, H. de (2012). Electroencephalogram signals classification for sleep-state decision–a Riemannian geometry approach. *IET signal processing* 6(4), 288–299.

Liu, Z., Sun, J., Zhang, Y., and Rolfe, P. (2016). Sleep staging from the EEG signal using multi-domain feature extraction. *Biomedical Signal Processing and Control* 30, 86–97.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering* 4(2), R1.

Matic, V., Cherian, P. J., Koolen, N., Ansari, A. H., Naulaers, G., Govaert, P., Van Huffel, S., De Vos, M., and Vanhatalo, S. (2015). Objective differentiation of neonatal EEG background grades using detrended fluctuation analysis. *Frontiers in human neuroscience* 9, 189.

mcgill (2018). *EEG*. URL: https://www.medicine.mcgill.ca/physio/vlab/biomed_signals/eeg_n.htm (visited on 08/07/2018).

Millan, J. d. R., Mourino, J., Babiloni, F., Cincotti, F., Varsta, M., and Heikkonen, J. (2000). Local neural classifier for EEG-based recognition of mental tasks. In: *ijcnn*. IEEE, 3632.

Moakher, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* 26(3), 735–747.

Motomura, S., Ohshima, M., and Zhong, N. (2015). Usability study of a simplified electroencephalograph as a health-care system. *Health information science and systems* 3(1), 4.

Mozaffarian (2015). Heart disease and stroke statistics-2015 update: a report from the American Heart Association (vol 131, pg e29, 2015). *Circulation* 131(24), E535–E535.

Ning, L., Jiang, X., and Georgiou, T. (2013). On the geometry of covariance matrices. *IEEE Signal Processing Letters* 20(8), 787–790.

Nuttall, A. H. (1976). *Multivariate linear predictive spectral analysis employing weighted forward and backward averaging: A generalization of Burg's algorithm*. Tech. rep. NAVAL UNDERWATER SYSTEMS CENTER NEW LONDON CT.

Obermaier, B., Guger, C., Neuper, C., and Pfurtscheller, G. (2001a). Hidden Markov models for online classification of single trial EEG data. *Pattern recognition letters* 22(12), 1299–1309.

Obermaier, B., Neuper, C., Guger, C., and Pfurtscheller, G. (2001b). Information transfer rate in a five-classes brain-computer interface. *IEEE Transactions on neural systems and rehabilitation engineering* 9(3), 283–288.

Omar, W. R. W., Fuad, N., Taib, M. N., Jailani, R., Isa, R. M., Mohamad, Z., and Sharif, Z. (2014a). Brainwave Classification for Acute Ischemic Stroke Group Level Using k-NN Technique. In: *Intelligent Systems, Modelling and Simulation (ISMS), 2014 5th International Conference on*. IEEE, 117–120.

Omar, W., Mohamad, Z., Taib, M., and Jailani, R. (2014b). ANN classification of ischemic stroke severity using EEG sub band relative power ration. In: *Systems, Process and Control (ICSPC), 2014 IEEE Conference on*. IEEE, 157–161.

Papoulis, A. (1977). *Signal analysis*. Vol. 191. McGraw-Hill New York.

Röschke, J. and Aldenhoff, J. (1992). A nonlinear approach to brain function: deterministic chaos and sleep EEG. *Sleep* 15(2), 95–101.

Sebestyen, G. S. (1962). Decision-making processes in pattern recognition.

Shao, S.-Y., Shen, K.-Q., Ong, C. J., Wilder-Smith, E. P., and Li, X.-P. (2009). Automatic EEG artifact removal: a weighted support vector machine approach with error correction. *IEEE Transactions on Biomedical Engineering* 56(2), 336–344.

Strand, O. (1977a). Multichannel complex maximum entropy (autoregressive) spectral analysis. *IEEE Transactions on Automatic Control* 22(4), 634–640. ISSN: 0018-9286.

Strand, O. (1977b). Multichannel complex maximum entropy (autoregressive) spectral analysis. *IEEE Transactions on Automatic Control* 22(4), 634–640.

Tang, Z., Li, C., and Sun, S. (2017). Single-trial EEG classification of motor imagery using deep convolutional neural networks. *Optik-International Journal for Light and Electron Optics* 130, 11–18.

Tuzel, O., Porikli, F., and Meer, P. (2008). Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (10), 1713–1727.

Van Trees, H. L. (2004). *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons.

Wang, Y., Zhang, Z., Li, Y., Gao, X., Gao, S., and Yang, F. (2004). BCI competition 2003-data set IV: an algorithm based on CSSD and FDA for classifying single-trial EEG. *IEEE Transactions on Biomedical Engineering* 51(6), 1081–1086.

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics* 15(2), 70–73.

(WHO), W. H. O. et al. *Global cancer rates could increase by 50% to 15 million by 2020 [Internet]. Geneva: WHO; 2018 [cited 2017 July 28]*.

Woestenburg, J., Verbaten, M., and Slangen, J. (1983). The removal of the eye-movement artifact from the EEG by regression analysis in the frequency domain. *Biological psychology* 16(1-2), 127–147.

Wong, K. M. (2004). *Notes on Detection and Estimation Theory*.

Wong, K. M., Zhang, J.-K., and Jiang, H. (2016). Multi-sensor signal processing on a PSD matrix manifold. In: *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2016 IEEE*. IEEE, 1–5.

Wong, K. M., Zhang, J.-K., Liang, J., and Jiang, H. (2017). Mean and Median of PSD Matrices on a Riemannian Manifold: Application to Detection of Narrow-Band Sonar Signals. *IEEE Transactions on Signal Processing* 65(24), 6536–6550.

Xu, L., Wong, K. M., Zhang, J.-K., Ciochina, D., and Pesavento, M. (2013). A Riemannian distance for robust downlink beamforming. In: *Signal Processing Advances in Wireless Communications (SPAWC), 2013 IEEE 14th Workshop on*. IEEE, 465–469.

Yoo, K.-S., Basa, T., and Lee, W.-H. (2007). Removal of eye blink artifacts from EEG signals based on cross-correlation. In: *Convergence Information Technology, 2007. International Conference on*. IEEE, 2005–2014.